THE ASSESSMENT OF PSYCHOLOGICAL QUALITIES BY VERBAL METHODS

P. E. VERNON

MEDICAL RESEARCH COUNCIL

INDUSTRIAL HEALTH RESEARCH BOARD

REPORT No. 83

THE ASSESSMENT OF PSYCHOLOGICAL QUALITIES BY VERBAL METHODS

A Survey of Attitude Tests, Rating Scales and Personality Questionnaires

By P. E. VERNON

LONDON: HER MAJESTY'S STATIONERY OFFICE

1938

(*Reprinted* 1952)

INDUSTRIAL HEALTH RESEARCH BOARD
1938

E. P. CATHCART, C.B.E., D.Sc., M.D., F.R.S. (Regius Professor Physiology, University of Glasgow, *Chairman*

F. C. BARTLETT, M.A., F.R.S. (Professor of Psychology in University of Cambridge)

R. COPPOCK (Secretary, The National Federation of Building Tra Operatives)

MRS. STUART HORNER, M.B., B.Sc. (H.M. Medical Inspector Factories, Home Office)

W. W. JAMESON, M.D., F.R.C.P. (Professor of Public Health the University of London and Dean of the London School Hygiene and Tropical Medicine)

R. E. Lane, M.B., M.R.C.P. (Medical Officer, The Chloride Electri Storage Co., Manchester)

R. K. LAW, M.P. (Member of the Medical Research Council)

B. A. MCSWINEY, M.B., Sc.D. (Professor of Physiology, St. Thoma Hospital, London)

SIR FREDERICK J. MARQUIS, B.Sc., M.A., J.P. (Joint Manag Director, Lewis's Ltd., Manchester)

J. A. NIXON, C.M.G., M.D., F.R.C.P. (Emeritus Professor Medicine, University of Bristol)

SIR DUNCAN WILSON, C.V.O., C.B.E. (H.M. Chief Inspector Factories, Home Office)

AIR VICE-MARSHAL SIR DAVID MUNRO, K.C.B., C.I.E., M.B., *Secre*

COMMITTEE ON INDUSTRIAL PSYCHOLOGY
1938

C. BURT, D.Sc., *Chairman*

F. C. BARTLETT, F.R.S.

C. S. MYERS, C.B.E., M.D., F.R.S.

SIR JOHN PARSONS, C.B.E., D.Sc., F.R.C.S., F.R.S.

T. H. PEAR, M.A.

SIR CHARLES SHERRINGTON, O.M., G.B.E., F.R.S.

P. E. VERNON, PH.D.

S. WYATT, D.Sc., M.Sc.

AIR VICE-MARSHAL SIR DAVID MUNRO, K.C.B., C.I.E., M.B., *Secr*

PREFACE

ecounts of the use and evaluation of psychological tests are ined in a number of reports previously issued by the Industrial h Research Board, the earliest being a review of the literature cational guidance by Muscio (No. 12), which may be regarded ancestor of the present publication. In these reports, the discussed have been for abilities, for skills, or for special nal qualities, e.g. those involved in "accident proneness"; stances, may be mentioned Nos. 31 and 53 on the use of rmance tests of intelligence in vocational guidance, Nos. 55 68 on tests for accident proneness, and No. 64 dealing with ional tests of dexterity.

he present report surveys critically tests of a different kind, attitude tests, rating scales, and personality questionnaires. value of these tests is admittedly less well-established than the of tests of abilities; yet the qualities they try to test are so rtant as to be worthy of every effort to put them on a tific basis.

hat the Board have realised the importance to industry of sing and endeavouring to estimate the influence on contentment fficiency of the mental attitudes and emotional traits of workers, own in several of the reports already issued by them, e.g. tigue and boredom in repetitive work (Nos. 56 and 77), and e nervous temperament (No. 61). The present report describes ttempts now being made to subject such personal characteristics curate measurement.

he field covered is wide; much of the work described, for nce, has been done in America. In this country, also, a wide est is taken in these problems, not only from their social and ational aspects, but owing to their importance in industry, e more and more attention is being paid to discovering and ng the temperamental qualities and abilities that influence an idual's adjustment to his work, and to exploring the attitudes nployees to their working conditions—and to their employers. he Board believe that this report, whilst of special interest to tigators doing field work in psychology and to students of trial and social psychological problems, is nevertheless relevant bstance to the interests and problems which the Board share industrialists. There exists, so far as they know, no collected ture which covers the ground to the same extent both graphically and critically.

May, 1938

906)

THE ASSESSMENT OF PSYCHOLOGICAL QUALITIES BY VERBAL METHODS

A Survey of Attitude Tests, Rating Scales and Personality Questionnaires

BY

P. E. VERNON, Ph.D.

CONTENTS

INTRODUCTION

The widespread realization of the importance of "the human factor" is a striking feature of present-day civilization. More and more attention is being paid to the instincts, sentiments, complexes and other psychological characteristics of human beings. In industry we attempt to discover the main temperamental qualities and abilities that influence an individual's adjustment to his job, and we explore the attitudes of employees to working conditions or to their employers. In education we try to guide parents and teachers as to the best means of dealing with children at home and at school, and treat the maladjusted and the delinquent at Psychological Clinics. Social psychologists even hope some day to be able to alleviate the world's political and economic ills by means of their studies of human nature.

Up till fairly recently the whole of our knowledge of people and their motives has been gathered by haphazard and unscientific methods, such as everyday observation, subjective inference and intuition. We are only beginning to formulate the general principles which should govern the collection of data and the deduction of conclusions, and are still more backward in the establishment of systematic laws which validly describe human thoughts and conduct. The number of impartial investigators in the social sciences is indeed rapidly increasing, yet their methods of attack on their problems are still highly fallible, when contrasted with the methods used for research in the physical sciences.

The claim that a scientific psychology or sociology can only be obtained by applying to human beings the experimental and logical techniques of the more advanced sciences would appear, to the present writer, fallacious. Nevertheless there is a tremendous need of controlled experimentation and exact measurement in our field. Such generalizations as we can make about human beings are still derived far too largely from uncontrolled observations and from interviews or discussions, or occasionally from the answers to widely distributed questionnaires. Such methods are probably essential in the present state of our knowledge for obtaining a general conspectus of any problem. Indeed it would be futile to attempt a psychological or sociological experiment in a factory, school, or home, unless we first possessed an intimate personal knowledge of the people and of the relevant conditions. Such personal knowledge is of course unscientific, although its trustworthiness may be greatly improved by adherence to such principles of interviewing and of case-study writing as are now available. But having obtained this general conspectus, we should certainly attempt to apply more refined techniques. It is with the description of one particular branch of these techniques that the present Report is concerned.

The specialized methods of the scientific psychologist are far too numerous to treat in a single Report, and we shall omit altogether what is perhaps the biggest class, namely the recording and measurement under controlled conditions of people's actions, or of the

products of their actions, and confine ourselves solely to verbal behaviour—either oral or, more usually, written. This will be further restricted to types of verbal behaviour which are susceptible to quantitative treatment, and so will exclude the ordinary interview, the clinical techniques of the medical psychologist, together with diaries or autobiographies and the like. Since there are already many excellent accounts of verbal tests of intelligence and of special aptitudes, we will deal only with the measurement of emotional characteristics. A more exact definition of our scope is :—a critical survey of the methods used by psychologists for obtaining records of the verbal attitudes and affective judgments of people either about themselves, about others, or about their environment, these judgments being restricted by the psychologists in such a way as to make quantitative treatment possible, and the resulting records being considered as indices or measures of certain quantitative psychological variables existing either in the persons who give the judgments or in the persons or items judged.

On the borderline of our field lie those written questionnaire techniques, where people supply their own answers, or check one of a series of possible answers, to questions dealing with their attitudes and sentiments. The questionnaires which we shall include or exclude may be illustrated by the following instance*. A question on pacifism was answered by 22,627 students in 70 American Universities ; 39 per cent. checked the answer that they would not participate in any war, 33 per cent. agreed that they would participate only in a war where their country was invaded; and 28 per cent. said that they would join in any war declared by the U.S. This technique and this type of result is not in itself a measure of a psychological variable, and so falls outside our scope. The question might however be expanded, or so combined with other questions as to constitute a test of pacifist sentiments, which would yield so many marks for pacifism either to different students or to different University groups. Such a test, which regards pacifism as something varying in amount, would be relevant to our survey.

Even the limited field which we have chosen is surprisingly wide and varied. So much work has been done with the techniques which we are to describe, mainly in America, that a great deal of selection is necessary. Only what appear to be the more important and more frequently employed tests will be mentioned, and it will often be impossible to present all the available evidence for the conclusions which are stated below. In general little stress is laid on the results which have been obtained from applications of these tests, ratings, etc., our primary object being to describe the methods and the kind of social-psychological material to which these methods are applicable. Perhaps the fullest sections are those that deal with the scaling of attitudes, etc. in equivalent units (§§ **11–14, 45–53, 87, 137**), and with multiple factor analysis (§§ **38, 60–70, 112–119, 144–**

* For a general outline of questionnaire techniques, cf. Vernon (1938).

152, 221–224, 235–237)†, the reason being that few general surveys of these techniques have been published so far, and none are readily available in this country. In contrast there are already numerous accounts of most of the tests and their chief results, e.g. Symonds (1931, 1934), Fryer (1931), Droba (1932), Jones and Burks (1936), Allport (1937), Murphy (1937), etc.

Although the Report concentrates on methods, yet it must not pretend to be a detailed practical handbook on how to apply these methods. Rather it is intended as a guide for showing what methods are available, where fuller descriptions of them may be obtained, and in particular what chief precautions are needed in using them.

An attempt has been made to draw on the work of British investigators wherever possible, and most of the illustrations in Chapter II are taken from such work. In subsequent chapters however the enormous majority of the studies cited are American. There cannot be the slightest doubt that the development of these methods is far more advanced in the United States than in any other country. References to the studies of German, French and other foreign psychologists will be conspicuous by their absence. This is not because of any lack of valuable contributions to the psychology of personality and of society in these countries—we need only mention the names of Freud, Jung, Adler, Spranger, Ch. Bühler, Lahy, Mira, Luria, etc. to belie the view—but solely because their employment of the quantitative techniques with which we are concerned is negligible. As the writer has shown elsewhere (1933b), one of the most pressing needs of contemporary psychology is the reconciliation and integration of the approaches characteristic of Continental and American investigators. The interpretative insight of the former should help to correct the blind empiricism of the latter; and the uncontrolled subjectivity of the one should be checked by the scientific discipline of the other. In point of fact considerable advances have been made in this direction during the past five years, and there is reason to hope for still further interpenetration in the future.

No one can hope to treat this field entirely without prejudice; his views on the psychology of personality will inevitably affect his judgments, particularly when, as in this Report, he attempts to go beyond mere description of the work, and to provide a general assessment of its significance. It is anticipated, moreover, that while some will disagree with the writer's criticisms of ratings, personality questionnaires, purely empirical tests, etc., others will regard the methods as still more fallible, and therefore more useless in practice, than he has done. To the former he would point out how meagre are the actual achievements of scientific methods in the study of personality, when contrasted with the amazing extent and variety of our non-scientific knowledge of, and ability to control, human

† These numbers refer to the sub-sections or paragraphs into which Chapters II-VII have been divided. References to the bibliography are given by means of dates.

beings in everyday life. The influence of experimental psychology and mental testing in the realm of educational and industrial *abilities* is immense ; but not in the realm of *emotions* and *motives*. There the main advances have come so far from clinical and theoretical rather than from experimental psychology. The latter group of objectors should be reminded once more that these methods are still in their infancy, and cannot be expected to be able to provide solutions to any and every social-psychological problem. They do not supplant, but supplement and check " common-sense " methods such as observation, interviews and ordinary questionnaires ; and when used with caution they can supply far more accurate and objective information upon a great variety of special problems than can the subjective and biased generalizations of which both " common-sense " and clinical psychology so frequently consist.

II.—GROUP SURVEYS OF ATTITUDES AND INTERESTS

A. Techniques for Collecting Group Opinions

1. We will begin with one of the simplest types of scale, which will serve to illustrate many of the techniques needed in the more elaborate tests ; namely, scales for assessing the relative preferences of a *group of persons* towards a *set of issues*. For example, Thurstone (1928), Bogardus (1933) and others have studied nationality preferences, i.e. the relative popularity of English, Scottish, German, Turkish, Negro and other peoples among American testees. Many studies have been made of the relative popularity of school subjects among pupils, two recent ones carried out in this country being Pritchard's (1935) and Shakespeare's (1936). Valentine (1934) and others have investigated the relative importance of the various motives which influence teachers in their choice of a career. A good instance in the field of industrial psychology is Wyatt and Langdon's (1937) recent study of the chief causes of satisfaction or dissatisfaction among factory workers, and of the types of employment most popular among women operatives. Newspaper competitions which determine the most popular film stars or novels, etc. of their readers, might also be mentioned.

For obtaining these various preferences, four main techniques may be distinguished ; simple voting, rating, ranking, and paired comparisons.

(i) *Voting*

2. The persons whose opinions are being studied are provided with a list of items (e.g. names of film stars, causes of dissatisfaction with work, etc.) and instructed to vote for the one item which each regards as most important, or for the two or three or more most important. The number of votes received by each item, or the percentage of judges who vote for each is then tabulated. It is preferable for each judge to record the same number of votes, otherwise those who, let us say, pick out six sources of dissatisfaction

will have a much greater influence on the final result than those who only choose one source. This difficulty may be overcome, however, if the former judges' votes are each given one sixth of the weight of the latters' ; and if the votes of other judges who give other numbers are similarly weighted.

This technique is the simplest from the point of view of the judge, but the final result is unlikely to be reliable unless a very large number of judges take part. For example, the relative importance of various causes of dissatisfaction found among one hundred operatives might be distinctly different from their relative importance among another hundred.

(ii) *Rating Methods*

3. Each judge is instructed to give, say, 4 marks to those items which he regards as important, 3 to items which are somewhat less important, and so on down to 0 for items which are to him of no importance. Eaglesham (1937) used this technique in finding the opinions of teachers on the relative practicability of a set of twenty-eight educational ideals. It has often been employed in industrial psychological investigations, such as rating the relative effectiveness of advertisements. Alternatively, plus and minus judgments may be requested, e.g. + 2 for strong likes, + 1 for moderate likes, 0 for indifference, —2 for strong dislikes ; or letters (A to E) may be substituted. Experiments by Symonds (1924) and others have shown that raters can fairly readily distinguish five or even seven different grades. A larger number than seven is unwise, because raters can hardly discriminate so many steps consistently, unless they are specially trained. A smaller number (e.g. +, 0, —, or A, B, C) is rather wasteful of the raters' discriminatory powers, just as is the voting technique. It therefore leads to decreased reliability of the ratings.

To combine the results of all the raters and determine the group preference, the ratings given to each item are summed and divided by the number of raters. If some of the raters have omitted some of the items, this averaging procedure will allow for it. Adjustments may, however, have to be made for two very common errors.

4. Errors in ratings and their correction.—First, some raters may adopt a much higher average rating than others ; e.g. A may chiefly give 4's, 3's and 2's, B chiefly gives 2's, 1's and 0's. Thus, when Thorndike (1935) asked a number of persons to estimate their degree of interest in, or liking for, various topics, on a + 5 to — 5 scale, he found that the average rating used by different judges ranged from + 4·2 to + 0·4.

Secondly, some raters may use more extreme ratings, either high or low, than others. In an experiment where the writer presented Eaglesham's list of educational ideals to 109 student teachers, he found that one student used 7 per cent. of 4's and 0's, another 59 per cent. instead of the desired 40 per cent. The judges were specifically instructed to employ approximately equal proportions of 4's, 3's,

2's, 1's and 0's ; but for this warning, the variation in dispersion of ratings might have been considerably greater.*

Now the first error is quite immaterial *so long as every rater rates every item.* But if, say, rater A gives high marks to items Nos. 1–23, and omits Nos. 24–28, while rater B gives low marks to Nos. 6–28 and omits Nos. 1–5 ; then in the final averages, items 1–5 will be too high, 24–28 too low. Hence if large errors of this type occur, the average of all the ratings given by each rater should be determined ; these averages should then be made identical by applying an appropriate correction, before the ratings for the separate items are summed : e.g. if A employs an average rating of $2\frac{1}{2}$ instead of 2, then half a mark may be subtracted from all his judgments.

The second error is less obvious but much more serious. If it is not corrected, rater A (with a large dispersion of ratings) will have a greater influence on the final averages than rater B (with a small dispersion). The appropriate adjustment is to express each rating as a " sigma score," i.e. as a deviation from the rater's own mean, divided by the standard deviation of all his ratings ; and only then to sum the results for the items. This procedure is simple, but laborious ; and it may have very little effect on the final results if the number of raters is large, and their variations in dispersion fairly small (cf. Conrad, (1933)). The investigator should perhaps first calculate the reliability of his results (cf. §**15**) without making these corrections, and only if the reliability is low need he attempt to see whether the application of the corrections will improve it.

A more regular distribution of ratings may be secured if the raters are instructed as to the proportions of items which they should assign to each step on the rating scale. It is usually assumed that the distribution should approximate to a " normal " or binomial type. For example, if the unit of the scale is taken to be the standard deviation, approximately 7 per cent. of the items would receive 4 marks, 24 per cent.—3, 38 per cent.—2, 24 per cent.—1, and 7 per cent.—0†. The appropriate proportions for other rating scales with different numbers of steps are tabulated by Symonds (1931) and Guilford (1936).

---

* Precisely the same errors occur in the marking of examination questions, when the marks given by different examiners, or the marks from different questions have to be combined. Both the average mark, and the dispersion of marks given by each examiner to each question are liable to scandalous variations, and should be adjusted (cf. Hartog and Rhodes, 1936).

† When the results of several raters are to be combined, their averaged ratings will show a much narrower distribution than that of the original ratings. The writer would suggest, therefore, that raters be asked to use a distribution with a large standard deviation. For instance, if 10 raters each adopt an 18, 20, 24, 20, 18 per cent. distribution (S.D. = 1·643), and if their ratings yield an average inter-correlation of +0·30, then the averaged ratings will yield the desired 7, 24, 38, 24, 7 per cent. distribution (S.D. = 1·0). Again, if the raters are likely to inter-correlate to +0·60, then a 12, 22, 32, 22, 12 per cent. distribution (S.D. = 1·25) should be used. If the average inter-correlation can be estimated, the appropriate distribution may be readily calculated from Kelley's formula No. 171 (cf. Kelley, 1923).

5. Obtaining constant standards among raters.—Another source of inaccuracy in this technique, which cannot easily be corrected, is that the rater's standards may vary indiscriminately throughout the rating process; e.g. he may give rather high marks to all the items at the beginning and later become stricter. (This may also occur in examination marking (cf. footnote p. 6.)) To reduce this we may adopt either of two procedures which are common in rating human character, namely the "Man to Man" scale (§ **84**) or the "Graphic" scale (§ **85**). The rater should be instructed to look through the items and pick out five of them which seem to him representative of the five grades which he intends to employ. Thereafter he should compare each item with these samples and give it the same mark as the sample to which it most nearly corresponds. Alternatively the investigator should supply the rater with a standard set of samples. This is commonly done in educational scales for grading the quality of children's handwriting or drawings, etc. Generally however it is impossible to select a set of samples upon which all the raters will agree.

6. Another way of reducing the difficulty and obtaining common standards among different raters may be illustrated by Bogardus's study of nationality preferences (1933). Instead of telling his judges to give 7 marks to those nations they liked most, 1 mark to those they liked least, or to use some other numerical scale, he instructed them to check one of the following statements with reference to members of each nation:

Would marry
Would have as regular friends
Would work beside in an office
Would have several families in my neighbourhood
Would have merely as speaking acquaintances
Would have live outside my neighbourhood
Would have live outside my country

The adoption of such a method seems likely to make the ratings more concrete and more reliable than any numerical method. Yet the ratings can of course later be translated into numbers by the experimenter.

7. The great advantage of the rating technique is its easy applicability to large numbers of items. The two following techniques are, for various reasons, more accurate than rating; but they can only be applied to relatively small numbers of items, No. (iii) to about 25 or less, No. (iv) to about 12 or less.

(iii) *Ranking*

Here each judge arranges in order of preference the series of items (nationalities, school subjects, advertisements, etc.). It is desirable to present the items on a set of cards so that they may be readily manipulated and re-arranged until the judge is satisfied with his order. To be given a printed list and to write 1, 2, 3, etc., in order of choice is less satisfactory, not merely because it is more

confusing, but also because there is likely to be a considerable "space error." Symonds (1936) has shown, both with children and with adults, that there is a tendency to rank items near the head of the list too high, and items near the foot too low.

When every judge ranks all the items, the combination of the results from all judges is simple. The average, or better, the median rank obtained by each item is determined. If desired, these final averages may be re-ranked.

This procedure does not, like rating, involve giving *absolute* marks to any of the items, but depends only on comparison of their *relative* desirability. Hence the two common errors in ratings (§ **4**) are avoided; both the mean and standard deviation of each judge's rankings are identical. Moreover, the judges' standards are not likely to vary during the process of ranking because they are, or should be, comparing every item with every other before deciding on their positions.

8. Distribution of rank orders.—We should note, however, that a rank order assumes a highly artificial distribution of items. With, say, 20 items, it conveys the impression that the difference between items 1 and 2, or 19 and 20, is the same size as the difference between Nos. 10 and 11; whereas it is highly probable that the latter items are closer together than the former. This is borne out by the difficulty which judges find in ranking the middle items to their satisfaction; they can usually differentiate much more readily at the extremes. Although we certainly cannot claim that all preferences and attitudes are distributed in the total population according to the normal curve (in fact Thurstone and Chave (1929) strongly criticize such an assumption), yet a normal distribution is generally more justifiable that the rectangular distribution yielded by ranks. For example, a typical group of school children might be very fond of a few school subjects, and very much dislike a few subjects, but be indifferent or undecided about the majority of subjects; and a group of operatives might strongly object to certain factory conditions, greatly appreciate others, but possess no attitude either way towards a large number of conditions. In both these instances, then, the distribution should approximate to normality. Hence there is much to be said for translating the results of a ranking study into sigma units, which do assume a normal distribution. This procedure may be facilitated by Symond's (1931) or Hull's (1928) tables. With 20 items, the sigma scores of items ranked:

1st 2nd 5th 10th 11th 16th 20th

would be:

$$+2{\cdot}06 + 1{\cdot}45 + 0{\cdot}76 + 0{\cdot}06 - 0{\cdot}06 - 0{\cdot}76 - 2{\cdot}06.$$

It will be seen that these scores *do* allow for the greater difference between 1st and 2nd than between 10th and 11th places.

9. Combining incomplete sets of ranks.—Sometimes it is impossible to get complete lists from all the judges; they may rank partial, but overlapping, sets of items: e.g. in Pritchard's (1935) investigation the school children from different schools dealt with different

lists of school subjects. In such cases it is usually adequate to sum the total ranks received by each item and divide by the number of judges who ranked it. Alternatively, Pritchard's method of averaging the deviations of each rank from the mean rank of its series may be fairer. Garrett (1924) has examined several other more exact methods, but fails to find any superiority in their results. For instance, the list supplied by each judge may be turned into percentile ranks, and the various percentiles obtained by an item then averaged ; or the lists may be turned into sigma scores before averaging.

(iv) *Paired Comparisons*

10. Here each judge compares each item directly with every other and records his preference. The pairs of items should be presented in random order. Thus if items *a*, *b*, *c*, *d* and *e* are to be judged, then the order might be *ac*, *db*, *ec*, *ba*, *de*, *cb*, *ad*, *be*, *cd*, *ea*. This technique was used by Thurstone (1928) in studying nationality preferences, by Wyatt (1937) in determining the main factors that influence satisfaction with work, and in innumerable other researches. If there are n items, the total judgments required from each judge $= \frac{n}{2}(n - 1)$. Hence the task becomes very laborious when n is large, and it does not, apparently, produce any better results than does the ranking technique (cf. Guilford, (1936) ; Saffir, (1937)). It can however be extended to larger numbers if the items are first roughly rated by the judges, and comparison of pairs is then applied to a few items at a time which are known (from the ratings) to be close together. (Thurstone (1930) has used the same device for extending the ranking method to large numbers of items.)

The final order of preference is obtained from the total number of times each item is preferred.

B. The Scaling of Group Judgments in Equivalent Units

11. All these techniques yield certain numerical values which represent the group's relative preferences for the various items. But such numbers suggest a spurious degree of accuracy. If, for instance, items *a*, *b*, *c*, and *d* happen to have received average ratings or ranks of 3, $2\frac{1}{2}$, 2 and $1\frac{1}{2}$, we might think that *a* is as much preferred to *b* as *b* is to *c*, and that *a* is twice as popular as *d*. But such a conclusion would be no more legitimate than the conclusion that a cake which costs 3*d*. is twice as nice as one that costs $1\frac{1}{2}d$. If our numbers were inches on a ruler or pounds weight, then the units would be equivalent and the conclusion would be true ; but we have as yet no knowledge of the size or the equivalency of our psychological units. Thus, although Wyatt's research shows that " Security of Employment " is the chief factor in job satisfaction for the average operative, yet we are not entitled to deduce from his figures how *much* more important it is than the other factors. Another defect of our present scales may be illustrated by the ranking of, say,

ten pictures in order of preference ; a certain average rank for each picture is obtained. Now suppose that an eleventh picture is added to the series and that they are all re-ranked, we shall indubitably find that all the previous ten averages have been altered slightly by this introduction of a new picture. If, however, our figures had consisted of truly equivalent units (e.g. if we had simply measured the linear dimensions of the pictures) then such figures would not be altered by inserting additional pictures into the series.

12. Comparisons of attitudes of groups.—One of the main uses of the techniques, which has not been mentioned so far, is the comparison of the opinions of different groups of judges. Shakespeare (1936) compared the school interests of children of different ages ; Eaglesham (1937) contrasted the educational ideals of Scottish and English teachers ; Wyatt studied work attitudes in different factories where different conditions prevailed. Wickman (1928), Stogdill (1934) and others obtained ratings from teachers, parents, and psychiatrists as to the relative seriousness or harmfulness to children's mental health of a number of bad habits ; the differences between their lists were very illuminating. Innumerable other such investigations might be cited. Now in none of the British experiments was it known precisely how significant were the differences between the groups, whether they might not be due to chance factors, inherent in the unreliability of the scales of preferences. The exact amount and the statistical significance of such differences can, however, be readily determined nowadays if the preferences are properly scaled.

13. Scaling of paired comparisons or ranked data.—The techniques used for scaling have been developed chiefly by Thurstone* on the basis of the psychophysical techniques used in grading sensation intensities and the like. He holds that, in so far as it is possible to assess degrees of brightness or loudness, and to say that the difference between the intensities of stimuli *a* and *b* is equal to the difference between *c* and *d*, to the same extent it is possible to assess degrees of popularity. For a full account of the theory and the application of the techniques the reader should consult Thurstone's articles (1927 *ab*, 1928, 1930, 1931*a*), or Guilford's book (1936) ; we will only attempt here a somewhat schematized outline of the method.

The principle employed is that equally often noticed differences are equal, that if the probability of *a* being judged higher in the scale than *b* is the same as the probability that *c* is judged higher than *d*, then the differences are psychologically equal. And the unit in terms of which these differences are expressed is the standard deviation of the distribution of preferences.

Knowing from our experimental data what is the probability that *a* is

* McK. Cattell, Thorndike and others had previously adapted psychophysical methods to educational scales and other psychological variables, e.g. for the purpose of grading a set of samples of handwriting according to merit in terms of equivalent units.

preferred to b, we can read off from the Kelley-Wood Normal Curve Tables the corresponding value of $\frac{x}{\Sigma}$; x then represents the true amount of the difference in rational units between a and b. We can perhaps clarify this principle, and show how Σ is obtained, by illustrating its application to the scaling of rank orders. Let N persons arrange n items, S_a, S_b, S_c.... S_n in order, assigning to them ranks R_1, R_2, R_3....R_n. Now if a considerable number of judges place S_a at, say, R_{10}, almost as many are likely to place it at R_9 and R_{11} ; rather fewer will be likely to place it at R_8 or R_{12}, and very few are likely to put it as high as R_5 or as low as R_{15}. The distribution of judgments tends then to conform to a normal distribution curve, centred around the mean value of the R's assigned to this S. Similarly the R's given to S_b will tend to be distributed normally around some other R, say R_{14}. These two distributions possess standard deviations σ_a and σ_b. Now it will be seen that the *differences* in rank positions given to S_a and S_b by different persons will also tend to vary. The commonest difference will be $R_{14} - R_{10}$ = 4 ranks, but among a few judges S_a and S_b may be as much as 10 ranks apart, and in a few S_b may be ranked higher than S_a. These differences then will also tend to be distributed normally, and it is this distribution of preferences which possesses the standard deviation Σ.

Now $\Sigma = \sqrt{\sigma_a^2 + \sigma_b^2 - 2r\sigma_a\sigma_b}$. Hence the required difference between S_a and $S_b = x = \left(\frac{x}{\Sigma}\right) \sqrt{\sigma_a^2 + \sigma_b^2 - 2r\sigma_a\sigma_b}$. $S_a - S_c$ or $S_g - S_d$, etc. may be similarly derived. In actual practice it is usual to determine the scale separations only for adjacent pairs of items, *i.e.* between pairs which are next to one another in average rank, since the greatest reliability is thereby obtained.

The calculation of $(\sigma_a^2 + \sigma_b^2 - 2r\sigma_a\sigma_b)$ for each pair would be extremely laborious, and would seldom be possible unless exceptionally comprehensive data had been collected. Thurstone (1927*a*) shows that we are usually justified in assuming no correlation between the preferences for a and b, *i.e.* in making $r = 0$, and that for most purposes it is adequate to regard σ_a, σ_b, σ_c, etc. as identical and equal to unity. In that case the calculations reduce to the simple quantity, $S_a - S_b = \frac{x}{\Sigma}\sqrt{2}$. For the scaling of ratings a somewhat different technique is employed, which is described in §§ **45-48**.

14. Applications of scaling.—We will give one illustration of Thurstone's results, namely, the scaling of criminals according to the seriousness of their crimes (1931*a*). Gangsters and kidnappers would generally be regarded as worse than bootleggers, who would be worse than pickpockets, who would be worse than tramps. By obtaining paired comparisons of 13 types of criminals from 240 children, Thurstone was able to quantify these differences. On a scale of 0 to 3 units, tramps are placed at 0·0, beggars at 0·2, drunkards and gamblers at 1·5, pickpockets at 1·9, bootleggers and smugglers at 2·6, gangsters and kidnappers at 3·0.

An important development was made possible by means of such scales, namely, the measurement of the effects of various types of propaganda. After sorting the criminals, the children attended a performance of the motion picture film " Street of Chance," which portrays gambling in an unfavourable light. They then sorted the criminals again, and on the new scale gamblers had gone up from 1.5 to 2.1 (a 20 per cent. rise), whilst the other criminals remained in almost identical positions. Similar experiments have been performed on attitudes to different races, to war and nationalism,

(40006)

to prohibition, etc. In the majority of these the propagandist film was found to have had positive effects, which sometimes lasted for a considerable period (cf. Peterson and Thurstone (1933)).

C. The Reliability of Group Attitude Surveys

15. Calculation of reliability.—Reliability may either mean the extent to which results remain unaltered if the testees or judges repeat their judgments after an interval (it being assumed that no propaganda or other influence has operated during the interval to affect their opinions) ; or it may mean the extent to which the results resemble the results obtained from another, similar, group of testees or judges. It is better to denote these as the repeat reliability and the consistency, respectively. With large numbers of items, repeat reliability is readily determined by inter-correlating the two sets of results ; with smaller numbers the average amount of change that has taken place in the second trial should be examined. For the determination of consistency the most suitable technique is to calculate the average inter-correlation between the judgments of the testees, by means of formulae which Kelley (1923) provides (No.171 for ratings, No. 172 for ranks). If we have N judges, and the average correlation between any pair of them is r, then we can predict by the Spearman-Brown formula that the correlation between the total results and the total results of another group of N judges will be

$$\frac{Nr}{1 + (N-1)r}.$$ *

This formula has another use. Supposing we find the reliability of the results to be inadequate, say, a coefficient of $+0{\cdot}60$, then we can predict from it that we should multiply the size of our group by six in order to get a satisfactory figure of $+0{\cdot}90$. Generally speaking, it is desirable to have at least a hundred judges in order to get a reliable set of ratings, rankings or paired comparisons.

16. Factors influencing reliability.—Reliability (consistency) depends not only on the size of the group of judges, but also on their homogeneity, and on the clarity or unambiguousness of the items. It will be reduced if the judges are very diverse in their opinions, and will be low if they find much difficulty in understanding the items, or are for other reasons uncertain about their preferences. Simple judgments of relative popularity are likely also to be more reliable than judgments of equivocal qualities such as the "effectiveness" of advertisements or the "practicability" of educational ideals, etc., unless these are thoroughly and concretely defined. If the quality to be assessed is composite, it is better to split it into a

* The legitimacy of applying the Spearman-Brown formula to attitude judgments might be questioned (cf. Guilford, (1936)). Remmers (1931), however, shows that its estimates, when applied to ratings of character traits, are fairly accurate, though not so accurate as with educational tests. And Rosander (1936) finds that increasing the number of sorters in scaling a Thurstone-type attitude test (cf. §§ **46-48**) increases the reliability of the sortings to the extent predicted by the formula. We therefore seem justified in using it in the present instance.

series of components which may be rated separately and then later combined (cf. analytic rating scales, § **90**). For instance, instead of merely trying to mark the " goodness " of a set of children's English compositions, it may be preferable to mark them separately for :

General impression
Mechanics (punctuation, spelling, grammar, etc.)
Content (range and appropriateness of ideas)
Style (expression of these ideas).

These four qualities may then be weighted according to their relative importance before they are recombined.*

Finally, the greater the diversity of items, the higher will be the reliability. It is much easier for judges to decide on their preferences when some of the children's compositions, film stars, advertisements, etc. are very good, some very bad, than when they are all mediocre.

17. Reliability and prejudice.—All the above factors which make for reliability tend to imply certainty or unanimity of opinion, and this may sometimes be interpreted as stereotyped prejudice among the judges. Thurstone (1928) points out that in a nationality preference test, scaled in equivalent units, the width of the scale or the extent of separation between the various nations (which is directly dependent on the reliability of the judgments) gives a measure of the univocality and bias of the judgments. Chant and Freedman (1934) applied the same scale to Canadian students and found a similar order of preference as among American students ; but the width of the scale was smaller, indicating that the Canadians were less unanimous in their likes and dislikes and more tolerant.

Much the same point is brought out by F. H. Allport's and Thouless's studies of the distributions of responses to single items. Allport (1934) showed that when there is pressure in a social group towards conformity to a certain opinion or a certain course of action, then the distribution of opinions and actions tends to be, not the usual normal curve, but a J-curve. For instance, the strength of opposition to contraceptive methods among Roman Catholics would yield such a distribution. And Thouless (1935) found a U-distribution of opinions on certain religious and other controversial issues. When a group of students was instructed to rate on a + 3 to — 3 scale their degree of certainty of belief or disbelief in such statements as : " There are such spiritual beings as angels," there were far more + 3 and — 3 judgments than intermediate judgments, such as as + 1, 0 and — 1, which would indicate uncertainty, or a tolerant, balanced opinion. On the other hand an issue which was free from ethical prejudice and from any pressure towards conformity (e.g. " Tigers are found in certain parts of China ") yielded an excess of intermediate judgments, that is, an ordinary normal distribution.

* In applying this analytic procedure for the purpose of improving reliability, we are almost certain to find what is described below (§ **106**) as the halo effect ; all the separate judgments will be much affected by the judges' generalized like or dislike for the compositions. Hence it would be unwise to take these qualities at their face value. But the effect is immaterial when they are going to be recombined.

Clearly these social forces or biases which influence responses to single items will also bring about high reliability of discrimination within a series of items.

D. The Validity of Group Attitude Surveys

18. By the validity of a test of, say, mechanical aptitude, we mean the extent to which the results of the test actually predict the testees' ability in mechanical work. No such objective criterion of validity is available in a test of character traits or affective attitudes. But we still wish to know what the results of the test or scale really show. Can we, for instance, accept Eaglesham's conclusion that Scottish teachers are less progressive than English, on the basis of their ratings of educational ideals; or Shakespeare's conclusion that children chiefly enjoy school subjects which involve activity or which embody concrete human interests; or that (from a newspaper competition of some years back) Fildes's "The Doctor" is the most popular picture known to the British public.

It seems probable that the validity of the type of scales which has concerned us so far is on the whole better than that of any of the scales or tests that we are to consider later. There are indeed weaknesses, but these are likely to be intensified in the instruments discussed below. A number of flaws which tend to influence the validity may be pointed out.

19. Representativeness of sampling.—In the first place it is very unsafe to make deductions about other groups, or about people in general, from results obtained with a particular group of judges (cf. Vernon (1938)). The popularity of "The Doctor" is indeed assured among those who entered that competition, but does not necessarily hold among all readers of that newspaper, and is most unlikely to apply among readers of *The Connoisseur*. Similarly, the *Literary Digest* straw vote before the 1936 U.S. presidential election did not validly predict the election result, though it did prove that the *Literary Digest's* method of sampling the population was totally inadequate (cf. Robinson (1937)). Clearly then it is essential to study the constituent members of the group thoroughly before applying their results to other groups. A complete study is unnecessary; only those characteristics which are likely to be correlated with the attitude that is being measured need to be controlled. For example, the churches to which members of the group belong need hardly be taken into account in studies of attitudes to factory work, but might have an effect on their nationality preferences. Socio-economic level, intelligence, sex, and possibly age are likely to correlate with almost all the attitudes we have mentioned. Thus, it is generally desirable to assess these (a fairly easy matter) and to take them into consideration before drawing any wide conclusions from an attitude survey.

20. Relation between verbal opinions and behaviour.—Equal caution is essential before assuming that the results of attitude tests will be predictive of actual behaviour. We would refer the reader

here to G. Allport's discussion of the theory of attitudes (1935), where he shows that attitudes should not be defined either as types of overt conduct, nor merely as verbally expressed opinions or beliefs, but rather as directing or motivating tendencies which lie behind both behaviour and opinion. Thurstone and Chave (1929) also consider that actions and verbal opinions may be equally fallible as manifestations of an attitude. Any piece of behaviour, or any expression of opinion, has so many determinants that tests and questionnaires can hardly be expected to cover them all. As Katz and Allport (1931) point out, our publicly exhibited attitudes (which is what the tests chiefly tap), our deeper and more private motives (which may sometimes be reached under favourable conditions), and our actions, are inter-connected in so complex an organization that many discrepancies between them may be apparent. For example, some persons may put Americans and French near the top of their lists of nationality preferences, and yet behave in a far from friendly manner towards particular American or French individuals. Again, many professed pacifists will doubtless enlist when war is declared. But these instances merely show that conflicting sentiments may co-exist, and that some attitudes may be flexible and open to suggestion. Thus, it would be unfair to criticize attitude surveys because they are incapable of revealing every side of human nature. Much more detailed study of the individual or group is necessary before we can predict accurately how he or they will feel and act in any specific concrete situation.

21. Dependence of validity on the judges' co-operation.—Next, the *Einstellung*, or attitude of the judges to the investigator and the investigation, is of prime importance. The great majority of people are quite incapable of understanding the object of a psychological or sociological experiment, and are only too likely to be suspicious, and therefore to produce only what they consider to be conventionally acceptable opinions, or to try to give the judgments which they think the investigator expects of them, or to refuse outright. Allport (1937) and Stagner (1933) point out that most of the successful attitude studies have been carried out on University students whose co-operation is readily obtained; applied to such groups as prisoners or farmers, the same scales may be entirely useless. The extent to which the results are influenced by these factors will, of course, depend largely on the type of information that is asked for. There will seldom be any embarrassment over producing opinions on film stars or advertisements, but there may be a good deal of hesitation over revealing attitudes to employment conditions, for fear lest they be passed on to the employers. Assurances of anonymity certainly help, and yet may fail to remove all the inhibitions against making one's private attitudes public. On the other hand, it may happen that the judges feel more secure when given a rating sheet or questionnaire to fill up than they do when asked the same questions in person, since they may be more doubtful of the anonymity of an interview.

An introductory talk to the group with the purpose of securing good co-operation and frankness is obviously desirable. Perhaps even better was Wyatt's method, where the judgments were obtained during an informal interview with each individual. For co-operation is essentially an individual affair ; different people need to be appealed to in different ways. A further disadvantage of applying the questionnaire to a group simultaneously is that neighbours are likely to be able to see one anothers' papers, and so some may be deterred from putting down unconventional opinions which their associates might ridicule. The writer has found this to be particularly prevalent among school children who are usually so crowded together that they can easily overlook one anothers' test blanks.

22. Mutual cancellation of variable errors.—We would repeat, however, that these methods are most often applied to issues about which there are few inhibitions, so that the validity of the results is much less affected that in some of the later tests. It should be remembered also that, as we are interested only in group results, a number of individual variations in attitude to the investigation will tend to cancel one another out. For instance, those judges who treat the whole thing flippantly may be counter-balanced by those who are over-conscientious. It is only the "constant errors," which are in the same direction in the majority of members of the group, that affect the validity ; other, variable, errors merely decrease the reliability. The same is true of errors that may be introduced by variations in the judges' temporary moods or by their recent chance experiences. When the writer asked his students for criticisms of a number of questionnaires and tests which they were answering, they frequently raised this point, asserting that they might answer differently on another day when they were feeling happy instead of depressed, or vice versa, or when they might perhaps have read something in a newspaper which influenced their judgments. Only if a large proportion was similarly influenced would it upset the group results.

23. Dependence of validity on judges' interpretation of the test.—A further important weakness in attitude surveys is our uncertainty as to the way in which the judges have interpreted the instructions and the test material. Although the objective conditions of testing may be the same for all, yet the same words may often have very different meanings for different people (especially when the words possess emotional content), different also from the meaning which the investigator intends. Again, however, we may claim that most of the tests and scales so far discussed are fairly unequivocal. Test items such as nationalities, school subjects, and the like, are concrete ; and the mental process of expressing a preference is very readily understood by all judges of reasonable intelligence. But the possibilities of misunderstanding are much more evident in more complex investigations such as Eaglesham's. Here the judges, teachers or students, were requested to rate a series of educational

aims (e.g. " To develop in the child the capacity for employing his leisure in fitting pursuits ") according to practical importance, or to the extent to which they would follow each aim in teaching 13-year-old elementary school children. Not only are the items ambiguous, but also the judgment of practicability is far more confusing than a judgment of like-dislike. Here then the validity of the results may be seriously questioned ; for it is conceivable that the Scottish teachers (who were found to be less progressive in their judgments than the English) were interpreting the instructions more literally and were actually rating the aims by the extent to which they did apply them in practice, whereas the English may, wittingly or unwittingly, have rated them more on the basis of desirability. Probably Eaglesham's conclusion is valid, since further lines of evidence point in the same direction, but we are not entitled to assume it from the original questionnaire results alone. As a general rule it is highly desirable to supplement the findings of any attitude survey with other data, if possible of a more objective nature. Although, as shown above (§ **20**), pieces of information about the overt behaviour of the judges are not in themselves adequate criteria of the judges' psychological attitudes, yet the validity of a subjective scale will be greatly enhanced if the available factual data do apparently fit in. To take another instance : a scale of popularity of film stars is unlikely to be much affected by errors of equivocality, and yet its validity will be improved if supplemented by box office data as to the drawing capacities of these stars.

24. With certain types of items it may be far from easy for the judges to express their real responses in terms of simple judgments of preference. They may wish to say : " I prefer A to B in some ways, and B to A in other ways," or, " It all depends on such and such circumstances." These difficulties reduce the reliability and make it especially necessary to find out what the judges did mean by their preferences. Particular caution is needed in the interpretation of omitted items and judgments of equal, or neutral (e.g. ratings of zero on a $+3$ to -3 scale). They may mean that the judge has never thought about that issue and has no opinion either way, or that he is unable to make up his mind in view of strong conflicting inclinations.

Some writers (e.g. Symonds (1931)) imply that such critical reactions and carefully thought out responses are unnecessary and undesirable, claiming that better results accrue when immediate affective impressions are given*. It may be that trained pyschologists can fairly readily adopt such a naive attitude while taking a

* The writer has been unable to discover in the literature any definite proof of this rather important principle. Some confirmatory evidence comes from Estes's (1937) study of the ability of various groups to judge personality. Artistically inclined persons were found to be greatly superior to psychologists and other University teachers, and the former claimed that they used intuitive, impressionistic methods of judging, whereas the latter were more analytic and intellectual. Further work, both in relation to attitudes and to judgments of character, is much needed.

test ; but they fail to realize that it does not come readily to members of other professions. In the writer's experience, objections to such items are very frequently raised by intelligent and cultured adults who have developed cautious and critical habits of thinking, and so dislike forcing their complex attitudes into the narrow categories which the experimenter provides. Other persons such as college freshmen and secondary school pupils, although high in intelligence, are somewhat less sophisticated, and so do not seem to find the same difficulties. Still others may, however, be too naive. Pre-adolescent children, for instance, may never have verbalized their attitudes, nor attained to the degree of abstract thinking necessary for expressing them in numerical terms, ranks, etc. (cf. Kelley and Krey, (1934)). They may indeed have strong feelings about the issues presented to them, but may be incapable of fitting their feelings into the verbal or numerical responses provided. Attitude tests and rating scales suitable for children can, and have been, devized ; but they need to be much simpler than those employed with adults of average or superior intelligence. Below a mental age of about 7 years it is, in the writer's experience, almost impossible to " get across " the simplest kind of group tests of likes and dislikes.

25. A danger among the naive which does not appear to have received the study it deserves is that, when their attitudes are relatively hazy and unverbalized, the test itself and the form of questioning may often provoke responses which would not normally have occurred to them. For instance, it has often been stated that school leavers possess vocational preferences which are entirely outside the range of their actual capacities. But a research in progress at the Cambridge Psychological Laboratory indicates that this tendency has been much exaggerated as a result of some of the tests employed ; that many boys may never actually have wanted to be " Architects, astronomers, consuls, surgeons, etc.," but that given a list on which they are asked to express their likes or dislikes for such occupations, they naturally put down " like ".

26. Precautions in devising and applying tests.—Clearly then the investigator who compiles the original instructions and test items needs considerable insight into the probable reactions of his judges, in order that he may get from them the most valid and reliable responses. Since he is himself quite likely to possess an affective attitude towards the issue in question, he needs to make sure that opinions contrary to his own are properly represented in the items given. Suppose, for example, an industrial employer had wished to carry out an investigation of job satisfaction similar to Wyatt's ; he might easily draw up a list of sources of satisfaction or dissatisfaction which to him seemed comprehensive, but which might omit several items that were of importance to his employees. Wyatt was able to avoid this pitfall since he had already made a thorough qualitative study of the main factors involved, and he was less likely to be biased than our hypothetical employer. Similarly Valentine (1934), before drawing up his list of reasons for entering

the teaching career, discussed the topic with a number of students and got them to suggest the main reasons. In general, therefore, the investigator should himself be familiar with all sides of the issue to be surveyed, and should if possible consult others who can provide him with fresh viewpoints on this issue. He should then try out a preliminary form of his scale with a small group, representative of the group whose opinions he wishes to survey, and ask for criticisms or difficulties or further suggestions, and only then begin the mass investigation.

27. Finally we would recommend the adoption of a hint from experimental psychology. In the typical laboratory experiment, the conditions are standardized, and conclusions are drawn from the observations by rigorous logical or mathematical procedure. But at the same time full introspections are obtained from the subjects in order to ensure that the meaning to them of the various conditions was also standardized, and to aid in the interpretation of the results. This same step is desirable in most attitude surveys; either the judgments should be discussed at an interview between the investigator and each judge, or, if time does not permit this, space should be left on the test blanks for spontaneous comments and for reasons why such and such items were preferred. The investigator cannot of course apply mathematico-logical processes in generalizing from such additional data; inevitably he selects and interprets, and so lays himself open to the charge of bias, though he can reduce this by including liberal quotations from the comments in his published results. Yet such comments help immensely to confirm or contradict the validity of the conclusions which he draws from the purely quantitative data. Good instances are provided by Wyatt's research on job satisfaction and Pritchard's on school subjects. Sometimes indeed the most interesting and psychologically important findings arise from this material, rather than from the attitude test itself.

So far we have dealt only with tests or scales for sampling group opinions. All the following tests can be similarly employed; but they are also supposed to be capable of measuring an individual's attitude, and therefore involve certain fresh considerations.

III.—TESTS AND SCALES FOR MEASURING ATTITUDES OF INDIVIDUALS

A. Description of Tests

28. The many sources of unreliability and the errors among individuals which, as we have seen, tend to cancel one another out in a group scale, make it desirable to adopt more elaborate methods in testing individuals. The responses of an individual to a test of nationality preferences might indeed suggest that he was, say, generally unfavourable to Fascist nations; or his choices in a questionnaire on "your favourite novel" might indicate his aesthetic taste. But such deductions would be extremely unreliable, and

they would entirely fail to *measure* his attitude, to show how unfavourable he was to Fascism, or how strong was his aesthetic sentiment. It might be thought that in order to determine a man's attitude to art, religion, etc. or his general sentiments such as pacifism, nationalism, tolerance, etc., we could simply ask him straight out. But again we should not be able to grade his answers quantitatively, to say how he compared in these respects with people in general. Moreover he would quite likely have only very vague notions as to what such a general attitude entails, and might interpret the question differently from our expectation ; hence the significance of his responses would be dubious. And direct questioning is obviously undesirable because he would be liable to produce merely conventional opinions on politics, morals, etc., which he regarded as appropriate to the particular social situation. Apart from conscious falsification, he might not know himself sufficiently well to be able to deliver a true judgment as to whether he was tolerant or prejudiced, artistic or Philistine, and so on.

29. Tests of radicalism—convervatism.—In general therefore an attitude test consists of a long series of questions, or of statements with which the testee is invited to agree or disagree ; and from the general trend of his responses, the amount or strength of his attitude may be deduced in quantitative terms. For instance, a so-called 'opinionaire' for measuring radicalism or conservatism might contain 20 to 30 statements, of which the following (from Lentz's *C–R Opinionaire*, (1934)) are typical.

The metric system of weights and measures should be adopted instead of our present system.

Even in an ideal world there should be protective tariffs.

Conscience is an infallible guide.

Armistice Day should be celebrated with less martial spirit.

After each statement is printed Yes, No ; or Yes, ?, No. (? meaning Doubtful) ; or True, False ; or +, 0, — ; one of these responses is to be underlined or encircled. Alternatively the extent of agreement may be denoted by ratings, e.g. + 2, + 1, 0, — 1, or — 2 ; or the 'multiple choice' technique may be adopted, where several possible responses are provided, one of which has to be checked. The following example is abbreviated from Vetter's questionnaire (1930).

What are your views on HEREDITARY WEALTH ?

(1) All wealth should revert to the State at death.

(2) Taxes should confiscate the bulk, leaving only enough for support of dependent women and children.

(3) Inheritances should be taxed on a rapidly graded sliding scale, up to about 50 per cent. for large inheritances.

(4) Very large fortunes should pay a reasonable inheritance tax, but not so high as to become confiscatory.

(5) Individual thrift and initiative should not be damped by any inheritance taxation.

Such statements are relatively concrete and specific, so that the testee's general attitude is likely to be expressed in the responses he chooses ; and a basis for grading the amount of his attitude is

provided by the proportion of his responses which point in the same direction, or by the strength of these responses.

30. The Allport-Vernon "Study of Values."—This test (1931) gives another illustration of the method. It is designed to measure an individual's relative standing on six main types of values or general interests: theoretical or scientific, economic or utilitarian, aesthetic or artistic, social or humanitarian, political or power-seeking, and religious or spiritual. The testee is required to rank his order of preference for the four answers to questions such as:

If you could influence the educational policies of the public schools of some city, would you undertake—
(*a*) to promote the study and the performance of drama,
(*b*) to develop co-operativeness and the spirit of service,
(*c*) to provide additional laboratory facilities,
(*d*) to promote school savings banks for education in thrift?

If he puts the answer (*c*) top, he scores 3 marks for theoretical values, and if he puts (*a*) bottom, he scores 0 for aesthetic values. The other two answers refer to social and economic values. By summing the answers to all the questions a range of marks of 0 to 60 is possible on each value.

31. Watson's test of fairmindedness.—A more elaborate disguise is adopted in G. B. Watson's ingenious test of fairmindedness or prejudice (1925). Obviously the name "prejudice" must not be mentioned, hence the testee's blank is headed "A Survey of Public Opinion." Nevertheless the degree of prejudice in his political and ethical opinions is deduced in six sub-tests. For instance, one of these contains a series of statements such as:

All Most Many Few No Roman Catholics are superstitious.

The testee who encircles either "All" or "No" is deemed to show prejudice in his response; any of the less extreme answers are accepted as a sign of tolerance. In another sub-test appear statements such as:—

In the United States 3 per cent. of the people own 60 per cent. of the wealth.

Following it are several possible conclusions that may be drawn, including:

The great incomes should be more heavily taxed.
Such a concentration of capital is inevitable if industry is to be effectively developed.
No conclusion here stated can fairly be drawn.

The testee who checks the first of these as a legitimate inference is credited with socialistic prejudice; the second—capitalistic; only if he checks the last does he obtain a mark for fairmindedness.

B. The Form of Attitude Test Statements

32. Research on attitudes is regarded in America as so important that much attention has been paid to the best type of questions or statements to use (cf. Stagner (1933); Droba (1932)). Kulp (1933)

shows that different tests have employed items which express:

(*a*) the testee's personal conduct,
(*b*) his beliefs or opinions,
(*c*) his judgments,
(*d*) actual facts.

Examples of these types of items might be:

(*a*) I intend to vote for the Socialist party at the next election.
(*b*) The country should vote socialist if it desires world peace.
(*c*) Socialism is based on sounder economic principles than is capitalism.
(*d*) A large proportion of the electorate votes for the Socialist party.

Kulp finds that even when identical issues are expressed in these various forms they may evoke somewhat different responses. The fourth form is clearly unsuitable; the facts may be regretted or applauded, but they are not genuinely debatable issues. Many writers prefer the first form; the testee's statements of his past, present or future actions seem likely to express his attitude most directly.

33. Wording of items.—Wang (1932b) gives a detailed list of rules for the wording of items. They should be short, simple and unambiguous. The following is bad, since it might be taken as representing opposition to, or support for, birth control.

Birth control legislation is a disgrace to our civilization.

Double-barrelled statements are always ineffective, since some testees may pay attention to one clause, others to the other (cf. §**46**). Most important of all is their relevance to the issue, and the extent to which they cover all sides of the issue. Kirkpatrick (1935) recommends that a careful analysis first be made of the content of the attitude, e.g. to find out what radicalism really implies, and that items should then be chosen to represent the products of the analysis each.

Rundquist and Sletto (1936) have demonstrated clearly that in any attitude test which asks for socially acceptable or unacceptable opinions, questions which state the acceptable viewpoint are not always answered in the same way as questions of apparently identical content which state the unacceptable viewpoint. For instance, if 40 per cent. of a certain group answers Yes or True to:

The Government's policy is subservient to big business interests

it will not usually be found that the same testees, or even the same proportion of testees will answer No or False to:

The Government's policy is independent of big business interests.

Slight differences in meaning cannot account for the different responses. It appears that there is a very general tendency for unpopular questions or items to arouse more emotion, and to be answered less rationally than the same items stated in the reverse way (cf. §**158**). Since the two types of question touch off different aspects of the attitude, it is recommended by these authors that both types should be included in a good attitude test, in equal numbers.

Another reason for including items which are " pro " and items which are " con " a particular issue, is that the testees are likely to read a mixed list more carefully. In an opinionaire which consisted solely of radically-toned statements, the testee would be liable to check Yes (or No) throughout, without properly considering their merits. The two types should, of course, be arranged in random order. Similarly in a multiple choice test the most " pro " response should sometimes be printed first, sometimes last. This is desirable also so as to eliminate " space errors." For Mathews (1927) has found that there is a considerable tendency in multiple choice tests to check the left-most, or top-most, responses rather than responses printed on the right, or at the bottom, of a series.

C. The Standardization of Attitude Test Items

34. The single test items quoted above are easy to criticize on the grounds that they would not necessarily show radicalism, scientific interests, prejudice, etc. But it should be remembered that a test consists of a considerable variety of such items, each one of which is only supposed to be a partial manifestation of the general attitude. Just as arithmetical capacity is measured not by a single sum but by a test including many varied arithmetical problems, so also is attitude measurement approached. And the total score on the test will be much more valid than the results of separate items. Moreover, every good attitude test is subjected to a thorough examination before publication in order to ensure that the items are consistent, also so as to prove that the test is reliable, and to provide norms.

As already mentioned, we have no criteria for proving directly the validity of our tests, though such indirect methods as can be applied are generally favourable (cf. §**59**). But we can and do try out the extent to which the items hang together consistently; thus we do not claim merely on the basis of our own *a priori* notions that the four statements from a radicalism-conservatism opinionaire, quoted above, are connected with radicalism or its reverse.

The two chief methods for investigating this matter are the method of internal consistency and the method of external judgments.

D. The Internal Consistency Method

35. The experimenter first compiles a much larger number of items than he will need in the final form of the test, and tries out this preliminary draft with a group of testees. He assumes that although many of the items may be poor, yet on the whole the errors among them are likely to cancel out, so that the total scores possess some validity. He then analyses each item to see whether or not the responses to it agree with the test as a whole, and calculates a statistical index representing its predictive value, i.e. the extent of this agreement. If the predictive index is too low, the item is either eliminated, or modified and tried out again until it proves

satisfactory. The best items are then used for the final form of the test*.

36. Item analysis techniques.—The various types of index, or techniques for comparing single items with the total test, are described by Lentz and Hirshstein (1932), Zubin (1934), and Guilford (1936). All of them, it should be noted, necessitate the use of a large group of testees; if the group numbers less than about two hundred, then the indices of the various items will be so unreliable as to be practically useless. Sletto (1936), for instance, found an average correlation of only + 0·35 between the predictive indices of a series of items which were derived from several different groups of moderate size.

(*a*) Tabulation of responses to the item given by those testees whose total scores were above and below the median, or those who were in the top and bottom quartiles or deciles according to the total test. The statistical significance of the difference in responses among the high and low groups may be calculated.

(*b*) Similar is the tabulation of the total scores of all those who gave each response to the item in question. This is less crude than (*a*), since it makes fuller use of the available data; but it is more troublesome.

(*c*) Determination of the correlation between the item and the total score. With items to which only two responses are provided, bi-serial *r* is employed.

(*d*) Lentz's method (for dichotomous items only) consists in summing the total number of conservative responses of those testees who gave a conservative response to that item and the total radical responses of those who answered it radically.

(*e*) None of the above methods allows for the difficulty that we wish to choose as items for the final scale not merely those that correlate highly with the test as a whole, but also those which do not correlate too highly with one another. To include two items to which all testees respond in the same way is obviously wasteful, however well they may agree with the total test. It is particularly desirable to have a wide variety of items so as to test as many different manifestations of the general attitude as possible. Items which inter-correlate highly are more likely to be homogeneous than varied. The calculation of all the inter-item correlations would be tedious, though feasible. Horst's (1934) method of "successive residuals"—which is too complex to describe here—is an alternative which satisfies these dual requirements fairly effectively.

37. Weakness of internal consistency methods.—(*f*) A grave weakness in all internal consistency methods is that they merely serve to select items which agree well with the original total score. The experimenter's preliminary draft may, for instance, embody two entirely dissimilar attitudes, which should have been measured separately. And yet, as Sletto (1936) has proved, the internal consistency method will pick out the items that correspond best to a composite of both these attitudes, and will fail to reveal the flaw.

* Humm and Wadsworth (1935), in their standardization of a personality self-rating scale (cf. § **142**), point out a possible flaw in this process. The consistency and validity of the items that are retained may quite possibly be affected by the omission of those that are rejected, since the response to any one item is unlikely to depend solely on the content of that item; it may be influenced to some extent by the testee's general set towards adjoining items in the test.

Other errors or biases may have been present in the original form and these will still persist in the final form. Thus, the experimenter may have wished to assess *general* radicalism; but his first set of items may have been predominantly political, failing to cover other aspects of the attitude; or they may mostly have been appropriate to a test of liberalism or tolerance rather than radicalism. The process of item analysis serves to increase rather than reduce such biases.

38. Application of factor analysis.—Many writers believe that an attitude test should measure only a single uni-dimensional variable, and that if the original test contains items expressing several tendencies, then it should be split up into its components.

Guilford (1936) and Whisler (1934) propose to apply the methods of factor analysis for this purpose. We shall discuss factorial methods below (§s **61-67**), and will merely point out here a flaw which seems to render Guilford's proposition impracticable, namely, that the factors extracted from a set of inter-correlated measures (or test items) depend on the particular set of measures or items with which the factorist started. Again, therefore, any bias which is inherent in his original selection of items will simply be repeated in the factors at which he arrives. For example, if his original selection contained no items bearing on aesthetic radicalism (i.e., the revolutionary attitude towards art), then it is most improbable that factorization would extract an æsthetic component out of the items which had been included. Whisler actually carried out Guilford's suggestion, and analyzed the correlations between the responses to 31 miscellaneous attitude questions. He did succeed in arriving at an apparently meaningful set of attitude factors (§ **66**); but it is obvious that these are resultants of the items with which he started, and that they can hardly lay claim to much significance because his original selection possessed no logical basis.

Experiments in the factorization of attitudes are certainly of value, since they will throw much light on those attitudes which are sufficiently simple to be scaled as uni-dimensional variables, or which overlap too closely to be worth measuring separately. But they are not yet in a position to show us which items to select and which to reject in our attempts to measure any particular attitude.

39. Conclusions.—In view of these difficulties, Kirkpatrick (1936), Allport (1935), and others including the present writer, consider that the insistence on strict uni-dimensionality is undesirable. An attitude should be accepted as a complex pattern of more or less inter-related tendencies, which may lose its significance if it is too much subdivided into simple variables. From this standpoint then the adequacy of the internal consistency method of standardization depends on the adequacy of the initial definition of the scope of the attitude, and the extent to which the original set of items covers this definition. In other words, statistical methods will assist greatly in refining the construction of a scale, but cannot produce a good scale unless there is also a sound logical basis behind the construction.

E. Scoring and Providing Norms for an Attitude Test of the Internal Consistency Type

40. Scoring and weighting of items.—In a test with dichotomous (Yes No, or True False) items, the score is generally the simple total

of items responded to in a certain direction, e.g. radically. But in order to eliminate the effect of omitted items, the total conservative responses may be subtracted from the total radical responses. Scores may then range from $+ n$ to $- n$, where n = the number of items. With triple items (Yes, ?, No), each item may score + 1, 0 and −1; 2, 1, and 0; or 1, ½ and 0.

Since some items will always be found, during the analysis, to have higher predictive indices than others, it would be quite feasible, not merely to eliminate the poor ones; but also to weight those that are retained proportionately to their predictive indices. This is seldom done, except perhaps in tests which contain a rather small number of items, since such weighting does not appear to improve the reliability to an appreciable extent. Moreover it greatly increases the labour of scoring.

41. Acceptability of items.—One other point to be examined in a test of this type is the degree of "acceptability" or "controversiality" of each item. An item which is answered by over 90 per cent. or less than 10 per cent. of a normal group of testees in the radical way is of very little use for differentiating people's radicalism. Usually the acceptability of all dichotomous items should be fairly close to 50 per cent.; i.e. they should be answered in one direction by about half of an average group. Thus an individual's final score is based more on the *number* of situations in which he expresses a certain sentiment than on the *strength* of the sentiment implied in the successive situations presented to him.

42. Scoring multiple choice items.—In a test where each item is provided with five or more grades of response, it is usual to assign numerals arbitrarily to the various responses, e.g. 1 for the most conservative, 5 for the most radical response. Likert (1932) tried out a more rational method, basing the scores on the *in*frequency with which a response was given, and turning these figures into the corresponding normal deviates or sigma scores. For instance, the five responses to one item were checked by 13, 43, 21, 13 and 10 per cent. of testees respectively; the corresponding sigma scores are − 1·63, − 0·43, + 0·43, + 0·99, + 1·76. He found however that the simpler method of scoring yielded almost as reliable a scale as did this more elaborate method.

In multiple choice tests the acceptability of the middle response of the five should be as close as possible to 50 per cent. among an average group of testees. The final score is, of course, based both on the number of items which the testee answers in a certain direction, and on the strength of his responses.

43. Norms.—Standardization of an attitude test involves the collection of norms. It would be useless to find that testee A gives radical responses to 40 out of 60 items unless we also knew the average and the distribution of scores among other members of A's social group. In obtaining norms the final form of the test must be applied to a large, representative, group of testees. Often it is desirable to establish separate norms for groups whose averages

may differ widely, e.g. for men and women, for students, for factory workers, etc.

F. The External Judgments Method

44. Many investigators, including Harper (1927), Symonds (1931), Vetter (1930), Lentz (1934), and Kirkpatrick (1935), have called in several judges or selectors in order to discover whether others besides themselves considered that all the test items were connected with the attitude they wished to measure. If, for instance, half a dozen competent people failed to agree that "Conscience is an infallible guide" really indicates conservatism, then that item would be eliminated. This seems to be a thoroughly sound principle, since it ensures that the attitude to be measured is fairly definite and uniformly identifiable by a group of judges. It might well be extended so as to provide a solution to the problem which we raised above, namely the definition of the scope of a complex attitude. For example the judges might decide what proportions of political, moral, aesthetic and other items should be included in a test for general radicalism.

45. Thurstone's scaling technique.—An ingenious elaboration of the external judgments method has been developed by Thurstone in his series of attitude scales. This enables him to measure attitudes in terms of equivalent units. We will outline the main steps in the construction of Thurstone and Chaves's scale for *Attitude to the Church* (1929). First a large number of heterogeneous opinions about the church was collected from various sources—editorials, conversations, and from statements written anonymously by students, etc., for example :—

My experience is that the church is hopelessly out of date.

I believe the church is the greatest institution in America today.

I like the ceremonies of my church but do not miss them much when I stay away.

I believe the church has a good influence on the lower and uneducated classes, but has no value for the upper, educated classes.

I am interested in a church that is beautiful and that emphasizes the æsthetic side of life.

By ranging far enough afield the investigator can cover all shades and degrees of opinion, and so escape the narrowness to which opinions thought out only by himself are liable. Opinions which are clearly irrelevant to the main issue should be dropped ; others which are insufficiently closely connected will be eliminated automatically in the subsequent standardization.

46. Having obtained 130 suitable statements, Thurstone and Chave applied to them the treatment which was described above in connection with attitude surveys (§ **3**) ; that is, they found out from a large group of judges how favourable to the church or how unfavourable was each statement. Paired comparisons or ranking could be employed, but usually a rating technique, based on the psychophysical method of equal appearing intervals, is preferred. (Saffir (1937) shows that all these techniques yield practically identical scales). Three hundred judges were each instructed to sort the

opinions into eleven piles, regardless of their personal views about the issue. The pile on the left (numbered 1) includes those opinions that each judge regards as most appreciative, that on the right (numbered 11) includes the most depreciatory ; the other piles are intermediate and should be approximately equally spaced out in their degrees of favourableness. The numbers of statements assigned to the piles need not be equal ; but no one pile should contain as many as one quarter of the total.

The distribution of positions assigned to each statement by all the judges is plotted, and a curve approximating to the normal type is usually obtained. The median position is found, i.e. the position above and below which 50 per cent. of the judgments of favourableness fall ; also the quartiles, and the semi-interquartile range, Q. The median position is taken to represent the scale value of each statement. If Q, the dispersion of judgments is very large, this indicates that the judges disagreed about its favourableness. Probably the statement is double-barrelled or irrelevant, and it should be dropped out.

47. The median positions or scale values of the first three statements quoted above were 9·1, 0·2* and 5·1, showing that they are very unfavourable, very favourable and intermediate, respectively. The fourth statement, presumably because of its double implication, had a high Q, namely 3·6 ; hence it was omitted from the final form of the test. The average Q of the statements retained was 1·67.

Next the statements were submitted to a process analogous to that used in internal consistency tests. Another large group checked those statements with which they personally agreed. Thurstone then calculated for each statement an "index of similarity," whose purpose is the same as the various predictive indices described above (§ **36**). The fifth statement quoted here had a poor index of similarity, proving that it was too irrelevant to the main issue ; hence it also was dropped. Finally, 45 statements were selected which passed both these criteria and which were well spaced out so as to cover the whole scale from 0 to 11.

48. The scale is now ready for use. The testee is instructed to check those statements with which he agrees. He does not usually confine himself to a single statement, but endorses a certain range of opinions (e.g. the majority of those with scale values lying between 6 and 9, or 2 and 4). His score is then the average or median scale value of his choices.

The difference between the method of scoring employed in internal consistency tests and in Thurstone's tests should be noted.

---

* A scale value will be less than 1 if less than 50 per cent. of the judges assign the statement to the second or higher piles, and if more than 50 per cent. put it in the first pile. The median position is then found by extrapolation, and Q is determined from the upper quartile and median instead of from both quartiles. The same procedure applies to statements placed by more than 50 per cent. in the eleventh pile.

In the former the items are of roughly equal "acceptability", and the score is derived from the "extensity" of the attitude. Here the items are strongly graded from low to high acceptability, and the score is based on "level" or "intensity". We might compare the two types to Thorndike's "width" and "altitude" approaches in the measurement of intelligence (1926).

As in internal consistency tests, so in these scales the favourable and unfavourable statements are mixed up on the test blank. And the testee is not informed as to their scale values until he has completed his endorsements.

49. Applications of Thurstone's technique.—This scaling technique has been very widely applied. Scales are available for measuring attitudes towards war, pacifism, communism, birth control, the movies, prohibition, evolution, negroes, and a number of other controversial issues. Uhrbrock (1934) has standardized a scale of fifty statements for testing favourableness among factory employees towards the management; the following are some examples, together with their scale values:—

(2·1) You've got to have a "pull" with certain people around here to get ahead.

(10·4) I think this company treats its employees better than any other company does.

(5·4) I believe accidents will happen no matter what you do about them.

It should be noted that these scales all deal with "pro" and "con" attitudes; but it should be quite possible to treat other variables similarly, provided that they can be sufficiently clearly defined for easy sorting, and that they can reasonably be regarded in terms of "more or less," i.e. as uni-dimensional. This might apply to radicalism-conservatism. Similarly Eaglesham's (1937) list of ideals in education might be sorted according to the extent to which they represent the progressive or the traditionalist school of educational thought. This done, testees (e.g. teachers) might check those ideals with which they agree, and so be scored for their progressiveness. There are, however, a number of more general attitudes which can be dealt with fairly effectively by internal consistency, but which would not be amenable to scaling. For instance it would hardly be possible to construct a graded series of statements with which to measure tolerance, or scientific-mindedness, or aesthetic values.

G. Criticisms of Thurstone's Technique

50. Reliability of sorting.—The sorting technique has been criticized by Rice (1930) on the grounds that judges with different personal opinions might give different judgments as to favourableness. However, experiments by Hinckley (1932), Ferguson (1935) and others prove that bias among the judges does not affect the score values. They would be almost identical whether the statements were sorted by ardent church-goers or by atheists. Ferguson's study also indicates that reliable score values may be derived from as few as twenty-five judges. Nevertheless, the technique is somewhat

laborious, and it has to be supplemented by the analysis of indices of similarity. Thurstone (1929) has described a method based wholly on the actual endorsements of testees, which eliminates the sorting. He calls this the "method of similar reactions." It does not appear, however, to have been widely applied.

51. Rationale of scaling.—Next, there is the objection which we raised above (§ **39**) to the narrowness of scales which insist on testing simple, uni-dimensional, attitude variables. Thurstone admits that attitudes are too complex to be fully described by a single figure on a linear scale. But he argues that it is quite legitimate to single out one aspect, such as favourableness-unfavourableness from the total pattern and measure it scientifically. For we do the same when we measure, say, the height of a table with a foot-rule; and we do not blame the foot-rule because it fails to give us a complete description of all the characteristics of the table. Moreover, he would claim that such simplification is justified by the resulting accuracy of his scales, and the equivalency of his units of measurement.

Kirkpatrick (1936), however, has disputed the validity of these units. They are clearly not analogous to physical units, both because they lack a true zero point, and because an attitude score is not (like a length of so many inches, or a weight of so many pounds) a multiple of some real unit. He, therefore, prefers the "extensive" type of measurement, where a score of say 40 does signify that the testee answered 40 unit questions in a certain way. Thurstone's scaling is nevertheless still generally regarded as superior in that the differences between scores of 4–3 and 3–2 are psychologically equal; whereas we are not entitled to claim any such equivalency for the difference between extensity scores of 40–30 and 30–20.

52. Reliability of scores.—Other writers (Allport (1935); Likert (1932, etc.)) have doubted whether such refined techniques as Thurstone's are appropriate for such rough conceptions as human attitudes. Likert finds the internal consistency type of scale to be quite as reliable as the Thurstone type, or more so. The reason for this is, of course, that the testee gives a graded response to a large number of items in the former, and merely checks a few items in the latter. The reliabilities of several Thurstone scales were found by Likert (1934) to be improved when the testees, instead of checking, marked every item 1 to 5 according to their degree of agreement. Some of the scales are doubtless better than others. When applying the *Attitude to War* scale, Miller (1934) noted that the average testee endorsed opinions which were very widely spaced apart, in fact covering two thirds of the total scale. This would imply very poor reliability. Probably, however, this issue does not rouse sufficiently definite and diverse opinions to be suitable for scaling.

In spite of their different derivations, it would seem that the two types of scale yield much the same results. Pintner (1933) found a correlation of + 0·78 between the Thurstone-Chave *Attitude to the*

Church scale and the Allport-Vernon measure of *Religious Values*, a figure which approximates to the reliabilities of the tests.

53. The conclusion which appears to follow from the above discussion is that for most purposes of attitude measurement, the internal consistency type of test is the simpler and more satisfactory, provided that it is based on a thorough analysis of the attitude in question, and that the external judgments of a number of competent persons besides the experimenter himself are employed in the preliminary planning and selection of items. But that for accurate research, e.g. on group differences, or on the modification of attitudes by propaganda or other influences, a scale standardized by Thurstone's technique is preferable.

H. Reliability and Validity of Individual Attitude Tests

54. In establishing repeat reliability, the scores of a group of testees who have answered the test twice are inter-correlated. Since, however, memories of the first application may affect the answers on the second occasion, it is better to compare two parallel forms of the test. More frequently the consistency is determined, by correlating the scores on one half of the test (e.g. the odd-numbered items) with scores on the other half (the even-numbered items), and then applying the Spearman-Brown correction. If this method is adopted with internal consistency tests, the scores of the group upon whom the item analysis was performed must on no account be used, since they will yield a spuriously high reliability coefficient; the test must be given to a fresh group.

55. Conditions affecting reliability.—The reliability of most attitude tests is quite high, coefficients of 0·75 to 0·90 or even more being commonly obtained (cf. Likert (1932); Lentz (1930, 1934); Thurstone and Chave (1929); Kirkpatrick (1935); etc.). The level of reliability is directly dependent on the heterogeneity or diversity of opinions among the individuals tested, not (as in group surveys) on their homogeneity. In studying the repeat reliability of his conservatism opinionaire, Lentz found that the average testee may change his responses to 15–20 per cent. of items; but that most of these neutralize one another, so that the total scores are little affected. Over long periods some attitudes seem liable to considerable changes; Farnsworth (1937), using Peterson's *Attitude to War* scales, obtained coefficients of 0·88 when Form B was taken a few days after Form A, but 0·30, 0·27 and 0·12 when taken 1, 2 and 3½ years later. By means of a technique which Thouless (1936) has described it is possible to determine how far unreliability is due to the inefficiency of the actual test, and how far to variations in the psychological trait which is being measured, i.e. to what Thouless calls function fluctuation. The writer has submitted Lentz's data to this technique, and finds the very high "index of function fluctuation" of 0·89. This means that such alterations as do occur on retesting are mainly due to personal alterations in the testees' degrees of

radicalism rather than to the unreliability of the test as a measuring instrument.

56. Relation between reliability and validity.—From experience gained in standardizing the *Study of Values*, the present writer has been led to the conclusion that it is a mistake to aim at too high a reliability in an attitude test, since it may be obtained at the expense of validity; a similar argument is put forward by Kirkpatrick (1936). Very high reliability is usually found when a test consists of homogeneous material which is repeated many times under identical conditions (e.g. simple reaction times). Probably therefore the more closely similar in content the items of an opinionaire, the higher its reliability. But if the items are very homogeneous and very numerous, the testee can hardly fail to realize that they all refer to his own radicalism, or other attitude, and he may therefore tend to answer each item more according to his personal opinion of his radicalism than in accordance with the way he really thinks or feels about that item. Now a valid test of radicalism is surely not one that, in effect, asks the testee fifty times over whether he regards himself as a radical, but one that presents fifty different situations in which various aspects of radicalism may be expressed. More experimental investigation is needed before this thesis can be established; and it probably applies less to some attitudes than to others. For instance, more direct questions which evoke the testee's own opinion of his attitude may be appropriate in testing attitude to the church, but would be fatal in the testing of educational progressiveness or aesthetic values. In order to measure the latter the questions must be heterogeneous and disguised as far as possible; and therefore the reliability should not be too high. Again, although reliability can always be raised by increasing the number of items (the Spearman-Brown formula seems to apply here), it is a mistake to make the test very long, not merely because the testees may get bored and careless, but also because they may approach the items in more stereotyped fashion and fail to consider each one on its own merits. A fruitful experimental investigation might be carried out in which, say, the results of short twenty-item attitude tests were compared with the results of the last twenty items of a long one-hundred-item test.

57. Factors influencing validity.—Individual attitude tests are open to the same weaknesses as group surveys (**§§ 20–25**), together with certain additional ones. The main difference is, of course, that the various errors which occur in individual testees do not now cancel one another out. If A is very inhibited, B an exhibitionist C flippant, D over-critical and cautious, the reliability will not suffer (as it did in group surveys), but the validity will. The testee's *Einstellung* towards the test situation and to the investigator must also be more carefully considered because the tests are often rather more personal, and there may be more hesitation over the disclosure of political, moral, and religious opinions. Vetter (1930) has demonstrated that the mere fact of answering a radicalism opinionaire along with a group of other testees, instead of answering

it privately, tends to make the opinions expressed somewhat more conservative.

58. Defects due to the form of the test.—The difficulties occasioned by the form of the test are much intensified ; for instead of ranking, rating or pairing his preferences, the testee has to express a large number of complex responses to complex questions in terms of Yes, No, or numbers, or a few restricted answers. To expect such answers to cover completely every individual's spontaneous reactions to the questions is obviously futile. However carefully constructed the questions and the provided responses, intelligent testees will wish to make innumerable qualifications, though, as mentioned above (§ **24**) the less sophisticated will be more amenable to this straight-jacketing. Yet the critical testee is not justified in concluding that the test is worthless because he cannot always give his natural response to each item. For it must be remembered that the score for an individual attitude (unlike the results of most group surveys) is based on a large number of items, and that the misrepresentations which he finds in particular items are on the whole likely to cancel one another out. Admittedly these difficulties may produce errors which tend to decrease the validity of certain items, but the errors are variable rather than constant, and so may have little effect on the final scores. This weakness is, of course, the penalty that must be paid if quantitative results are to be obtained ; it would be quite impossible to measure attitudes or to compare individuals objectively if each individual expressed his attitudes in his own way. Nevertheless, the more care that is given to the compilation of items and to the preliminary trying out and discussion, so as to cover various shades of opinion as completely as possible, the better will be the test. And it would seem well worth while leaving space for spontaneous comments at the bottom of the test blank, or discussing the items with the testee after he has filled in his answers, in order to obtain more insight into his attitude to the test and into the significance of his score. That the forcing of responses into simple categories does not greatly distort the attitudes was well demonstrated by Stouffer's experiment (1931). He obtained from 238 students anonymous accounts of their own experiences and opinions about alcohol and prohibition. These completely unforced expressions of attitude were rated by four independent judges as to their favourableness or unfavourableness to prohibition. The ratings were then found to correlate + 0·81 with the scores of the same students on a test of *Attitude to Prohibition*, the (split half) reliabilities of the ratings and the test being respectively 0·96 and 0·94.

59. Empirical evidence of the validity of attitude tests.—We may claim then that when reasonable precautions are taken in constructing the test, and in obtaining co-operation from the testees, the validity should be good. And there is a large amount of scattered evidence supporting this conclusion. Moderate correlations are obtained between testees' scores on the *Study of Values* and

estimates by their friends of these testees' aesthetic, economic and other interests (cf. Vernon and Allport (1931)). Thurstone quotes similar results with his scales. We would not, of course, expect to get perfect coincidence between the attitude which the individuals express in the test and the attitudes with which their friends credit them. Again Cantril (1932, 1933) has shown that the results of the *Values* test possess a wide predictive value ; e.g. they correlate with speed of free word association to aesthetic, economic, etc. word lists, with the topics in newspapers that people with different interests notice when they scan the headlines, and so on.

Both this test and a great number of attitude scales have been found to differentiate effectively between groups of persons who might be expected to possess contrasting attitudes. Women consistently come out higher than men in aesthetic, social and religious values, lower in scientific, economic and power-seeking. Business, theological, science and art students, etc., obtain appropriate results. On Thurstone's *Church* scale the average score of Roman Catholic students was 2·90, of Jewish students 5·44. Uhrbrock (1934) finds significant differences between the attitudes to employers of foremen, clerical staff and operatives. On Watson's test (1925), social psychology students in an eastern American University are much more fairminded than Middle-West parsons. Vetter (1930), G. Allport (1929) and others note meaningful relations between radical attitudes and income of the parents, Jewish race, etc.; Klein (1925) connects up the attitude with antagonism of the testees to their own fathers. Likert (1932) shows that University students in the southern parts of U.S.A. are much more opposed to negroes than those in north-eastern parts. G. Allport (1929), Watson (1929) and others find that increased knowledge about an issue correlates, sometimes quite highly, with a more progressive, liberal or tolerant attitude towards that issue. Many of the scales have been used successfully in investigations of the effect on attitudes of propaganda or instruction. Not all of these lead to positive results. For instance, the mere study of a scientific subject at school or college fails to " transfer " or to produce a scientific attitude towards ethical and political issues. A full account of these researches is given by Lichtenstein (1934).

Thus, although no satisfactory objective criterion of attitude is available, yet such criteria as we possess seem unanimously to indicate that these tests do measure something significant, which does express itself in a wide variety of practical situations, as well as in verbal opinions.

I. Indirect Measures Derived From Attitude Tests

(i) *The Interlocking of Attitudes : Factor Analysis*

60. The facts which we have just mentioned as to the widespread manifestations of an attitude fit in with the view (cf. § **39**) that the attitudes dealt with by psychologists are not simple, discrete variables, but complex, overlapping structures. Particularly striking

inter-relationships were revealed in Katz and Allport's (1931) study (by means of a questionnaire) of the opinions on various topics of some four thousand students. For example, fraternity members differed from other students not only in attitudes to college affairs, but also in political, racial and religious opinions. The following instance is taken from an investigation by Zimmermann (1934). Testees were classified into two groups one of which regarded God as " a personal being " or as " a power making for righteousness," and the other regarded God as " a projection of our social conscience " or as " the universe as a whole." In other questions it was found that 71 and 42 per cent. respectively of these groups were in favour of prohibition, 13 and 26 per cent. in favour of socialism, 42 and 65 per cent. for birth control, and so on. Apparently some general tendency or tendencies are running through the various specific responses. These particular results were derived from answers to questionnaires ; but much the same is found if a series of standardized attitude scales are inter-correlated. Significantly large coefficients are almost always obtained. For instance, in Carlson's (1934) investigation, all the correlations between communist, pacifist, atheist, and anti-prohibitionist attitudes were positive.

61. Object of factor analysis.—It would then be a considerable advance if, instead of recognizing perhaps a thousand or more attitudes, interests, sentiments and ideals among human beings, and attempting the impossible task of measuring each of them in turn, we could generalize and abstract a few distinctive and fundamental tendencies in terms of which all the attitudes could be classified. There might conceivably be quite a small number of basic elements, somewhat analogous to the chemical elements ; and the attitudes with which we are dealing at the moment might be analysed, like chemical compounds, into combinations of these elements. Now there have developed recently certain statistical methods which claim to be able to accomplish precisely this analysis of a set of psychological variables into their underlying components. These originated in the work of Spearman (1927) on the fundamental factors behind intelligence and other abilities (work which has been described in Report No. 53 by Earle and Milner (1929). They have since been greatly elaborated by Kelley, Thurstone, Hotelling, Holzinger, and other statisticians in America, and by Burt, Thomson and Stephenson in this country. It is impossible to describe here these highly technical methods, but we will endeavour to outline their aims and achievements and, later, their limitations.

62. Spearman's technique.—All start out from a table of the correlations* between a series of tests or psychological measurements, which have been applied to the same group of testees. Now a correlation between a pair of tests implies that there is something common to them. Spearman found, when working with correlations

* Burt (1937*a*) shows that it would be preferable to factorize co-variances rather than correlations, and by so doing he is able to simplify Hotelling's method.

between tests of abilities, that all the correlations could be accounted for by the assumption of a single common factor which he called *g*—the general intellectual factor. The correlations were, of course, always less than unity, and this was because each test was considered to depend on a specific factor, *s*, found in that test alone, in addition to its dependence on, or "saturation" with *g*. What is known as the "tetrad difference" technique was developed for dealing with data of this type; it is easy to use, though very laborious. We cannot however expect psychological tests always to yield so simple a picture; there may often be more than one common factor running through them. Spearman's technique is less well adapted for dealing with such data. Holzinger's (1937) extension of it, which he calls the "bi-factor method" is claimed to be capable of isolating factors additional to *g*; but full details of this are not yet available*. Kelley's (1928) elaborate method depending on the "pentad criterion" had the same aim, but it has now been discarded in favour of the newer *multiple factor analysis* techniques. Of these Thurstone's (1933, 1935) simplified "centre of gravity" method is perhaps the most lucid and systematic, and is the most widely used at the present time.

63. Thurstone's technique as applied to student teachers' abilities. We will illustrate this technique from results obtained by the present writer in analyzing the marks of some 360 students at a Training College. Correlations were calculated between marks in teaching skill, speech training, education, psychology, hygiene, English, arithmetic, physical exercises, and a general intelligence test. The highest correlations were between education, psychology and hygiene, that is the theoretical subjects. By Thurstone's technique a general factor was extracted which accounted for most of these correlations; it was found to enter most prominently into the theoretical subjects, to a somewhat lesser extent into English and arithmetic, less still into speech training, teaching skill and physical exercises. Clearly it corresponded roughly to a general aptitude for the theoretical aspects of teaching. The influence of this factor on the correlations was removed and when the residual coefficients were examined, a second, entirely independent factor was obtained, which was highly loaded with the three practical subjects, but which showed no relation, or a negative relation, to the others. This also was removed and a third, smaller, factor was obtained which agreed positively with the scientific and negatively with the literary subjects. These factors are, as it were, three dimensions in terms of which various teaching abilities may be classified. In many investigations the analysis is carried further, and four, five, or more factors are extracted, in an attempt to account more completely for the observed correlations; but such supplementary dimensions tend to be rather inaccurate, and of doubtful value.

The process of factorization does not usually end here, since the obtained dimensions are often somewhat impure, and difficult to identify. The next step is to portray the results graphically. Putting the first (theoretical) factor as the X-axis, and the second (practical) dimension as the Y-axis, we plot a point for each set of marks corresponding to its factor loadings, i.e., its saturation with the two factors. (In this instance a three-dimensional

* This method, we now learn, assumes that each test can be regarded as involving one group factor (i.e. a factor common to a few, but not to all, of the tests), in addition to *g* and *s*. In investigations of abilities this assumption is usually justified; and as the method is simple and accurate, it should prove very useful. But for personality test data, multiple factors appear to be needed.

graph was needed, hence the points were plotted on the surface of a sphere). Thurstone has shown that more expressive factors may be obtained by choosing a fresh set of axes, passing through the same origin as the initial ones, but rotated through a small angle. The choice of the most suitable axes is guided by logical consideration of the inter-locking tests ; and its aim is to produce as many high positive and as many zero loadings as possible in each factor, in place of many moderately high positive or negative loadings. When the axes have been decided, the new co-ordinates of the points on the graph give us the new factor loadings of the tests.

The three new dimensions, so obtained, afforded a more logical classification of student teachers' abilities than did the initial ones. They could readily be identified as : (1) General teaching ability (2) Scientific-logical ability (3) Literary-humanistic ability. The full results will be published elsewhere, but one or two examples of the " saturation co-efficients " of the original subjects will illustrate their usefulness. The loadings of speech training with the three factors were + ·66, ·00 and + ·24, respectively ; i.e. this subject is very important in teaching, depends also on literary, but not on scientific ability. Arithmetic similarly gave + ·20, + ·56 and ·00 ; the intelligence test ·00, + ·30 and + ·28. The latter, as has often been found, does not affect teaching ability, but is moderately related both to scientific and to literary abilities.

64. Hotelling's, Kelley's and Burt's techniques.—These techniques are more mathematically perfect than Thurstone's, since they account completely for all the test inter-correlations. But they seem at present to be less useful to the applied psychologist, in that they are more complicated and that it is more difficult to interpret the psychological significance of their factors. One advantage of Hotelling's (1933) method over Thurstone's is that a testee's scores on each of the extracted factors can be calculated more readily (cf. Flanagan, 1935). Kelley (1935) asserts that the bases of his or Hotelling's methods are irreconcilably different from Thurstone's. But Burt (1937a) shows that they yield much the same results, at least in respect of the major factors, only by different routes ; indeed the products of a Thurstone analysis may be regarded as first approximations to the ideal solutions provided by Hotelling's " iterative " or Kelley's " trigonometrical " methods. Burt's own method of " higher moments " (cf. Hartog, Rhodes and Burt, 1936) gives these same ideal solutions, but with a considerable reduction in mathematical labour.

65. Applications of factorization : the general radicalism factor.—So far few factorial studies have been carried out with attitude tests. Kulp and Davidson (1934) applied a Spearman analysis to brief opinionaires dealing with racial, imperialist, international and other attitudes, and found good evidence of a general factor running through all of them, which they identified with liberalism. Thurstone (1934) applied his method to the correlations between eleven of his scales and obtained a conspicuous radical-conservative common factor, and a second smaller factor which seemed to represent, chiefly, a nationalistic tendency. Carlson (1934), Lentz (1934) and others have also noted fairly high correlations between different tests suggesting a common radical or progressivist tendency. The unanimity of these studies is rather striking in view of the many doubts that have been expressed as to the existence of a general

radical-conservative tendency (cf. § **74**). We might for instance suppose that people could be left wing in politics, conservative about morals, and undecided in other spheres. Some are no doubt relatively specific in their inclinations, but apparently a fairly consistent attitude in all spheres is more common*. The factor also correlates positively with educational attainment and, to a small extent, with intelligence (thus Thurstone (1934) finds, in a student group, a correlation of —0·44 between an intelligence test and a scale for patriotism). So far this is the only dimension of attitudes and interests which can definitely claim to be established.

66. Factorial studies by Whisler and Lurie.—We have already mentioned (§ **38**) Whisler's (1934) analysis of 31 general attitude questions. Six factors were extracted which he identified as follows :—

I. Acceptance of conventional ethical standards
II. Enjoyment of fleeting pleasures (or youth v. age)
III. Interest in conflicts and controversies
IV. Interest in controlling people and manipulating things (or sense of power)
V. Interest in social participation
VI. Sophisticated, critical attitude.

As shown above, the significance of these is dubious, since they depend so much on the particular test questions which the author happened to select. More useful results seem to emerge from an analysis by Lurie (1937) of the six types of general interest which the *Study of Values* (§ **30**) attempts to measure. Here the selection of test questions was based on a logical scheme of classification which had been very thoroughly worked out by the German philosopher E. Spranger (1928). Four sets of six questions referring to each value, 24 sets in all, were factorized ; after rotation of axes, factors corresponding rather closely to Spranger's types were obtained, namely :—

I. Social and altruistic attitude (composed largely, though not wholly, of items designed for testing the social-humanitarian type)
II. Philistine attitude (this combined items from three types, economic power-seeking and, negatively, aesthetic items)
III. Theoretical and scientific attitude
IV. Religious and spiritual (as opposed to economic-utilitarian) attitude

Three more factors of lesser importance included an open-mindedness or liberal *v.* conservative tendency, and a practical or materialist tendency. Lurie concludes : " It is believed . . . . that a more plausible and self-consistent system of personality classification can be founded on the four types derived by factor analysis than on the six types which Spranger developed by intuitive analysis of experience." Although the present writer would hesitate to accept this view until further evidence is available, he would admit that the

* The generality of the radical-conservative tendency is possibly greater in the American than in other cultures. The writer has, however, obtained evidence of a similar phenomenon among Scottish students, the more ardent church-goers being on the whole more conservative politically. It is also likely to be much more prominent at the student age than in more mature adulthood.

statistical investigation is a valuable corollary to the logical analysis, and that it helps to show which attitudes are, or are not, sufficiently distinctive to merit separate measurement.

67. Difficulties in the identification of factors.—We will return to a critique of factor analysis in Chapter VII, when we have seen the results that it yields with tests other than attitude scales. But we must note here one of its weaknesses, namely the difficulty in interpreting the psychological meaning of the factors. Almost the only way we have of identifying them is to examine which tests or test items are most highly saturated with them (this was done, above, in factorizing the student teachers' marks). Naturally we must not expect a factor to correspond precisely to any one conception or trait with which we are familiar in everyday descriptions of human beings. It is likely to cut across such conceptions, since it is in essence an abstraction from tests of a number of different traits. For instance, Thurstone was doubtful whether the common factor in his attitude scales (§ **65**) represented primarily a radical or an anti-religious tendency; possibly it should be interpreted as a blend of the two. Not infrequently a factor appears to involve an assemblage of traits so meaningless that one finds great difficulty in believing that it really corresponds to any basic personality tendency. The following is a rather extreme example, namely, Whisler's fifth factor which he considered to involve mainly social participation; actually it is composed of the following cluster of test items :—

Rarely thinks about the meaning of life
Does not enjoy spicy and highly seasoned foods
Much interested at a play in whether the characters violate conventional codes of behaviour
Prefers working with people to working with things or materials
Considerably interested in politics
Daydreaming has increased during the past five years
Differs considerably from intimate friends in interests, etc.

Copeland (1935) has criticized the laxness of the psychological side of factor analysis, contrasting it with the meticulousness of the mathematical side. He suggests that decisions as to the meaning of the factors should be based upon group judgments instead of merely upon the investigator's personal interpretation.

(ii) *Factor Analysis Applied to Correlations between Persons*

68. Burt (1937b) has recently pointed out an alternative approach to correlational and factor studies, based on the resemblance between persons instead of between tests. Supposing that fifty persons each take fifty tests, it would be just as easy to calculate the correlations between the scores of each pair of persons on all the tests as to perform the more usual calculation of correlations between all the testees' scores on each pair of tests. And the resulting 1225 correlations might equally well be submitted to Thurstone's or some other multiple factorization technique. Stephenson (1936 a, b, c, d) calls this the "Q-technique," or "inverted factor analysis" (the latter

title is, however, somewhat misleading). He has carried out by means of it a number of interesting studies of personality.

It is somewhat difficult at first to grasp the implications of this approach. But consider a pair of testees who are found to inter-correlate very highly; they resemble one another closely because they obtain much the same relative scores on the fifty tests. And if a common factor is found which will account for a number of the inter-person correlations, this will mean that all these persons approximate to a common type, since they all have similar relative scores. Again, just as in a straightforward factorization of tests, some tests are found to be highly saturated, some poorly saturated with a factor; so in this "inverted" factorization, some persons will be found to approximate to the type more closely than others. If the analysis is continued a second type (i.e. factor) may emerge to which a number of other persons approximate—persons whose test scores are all similar but are all dissimilar from the scores of the first type. Further types may be similarly extracted, but these are likely to fit smaller numbers of persons; just as in ordinary factor analysis the third, fourth, and later factors are generally of minor importance and encompass fewer tests. We have seen that test-factors may be lacking in universality, since they depend on the particular tests which are analyzed; in the same way these person-factors or types may be functions of the particular persons studied, and may fail to hold among different sets of persons. Either then a representative selection of persons, or a group chosen on some logical plan, should be factorized.

69. Such factorizations of persons are shown by Burt to be complementary to the straightforward factorization of tests. Indeed the scores obtained by an individual on a set of test-factors should be identical with his loadings in respect of a set of types or person-factors. Hence the choice as to which of the two approaches should be used depends simply on the kind of test scores or psychological material that is available.

For the determination of a correlation between persons, each person's test scores must be expressed in the same units of measurement, and must be approximately normally distributed within the person. Hence this approach is most useful for dealing with self-ratings or self-rankings on a series of traits (**§§ 151, 152**), or with assessments by each person of his own preferences for a series of issues such as school subjects, pictures, etc. (**§ 70**). It is less appropriate for treating the results of objective tests for the same reason, and because of the difficulty of persuading even a small number of people to submit to fifty tests which are sufficiently extensive to yield reliable scores. (A considerably larger number than fifty would be desirable, just as more than fifty testees are usually employed in finding the correlation between a pair of tests).

70. Applications of "Q-technique."—A study of children's interests in school subjects is reported by Stephenson (1936d). Twenty boys and twenty girls rated their preferences for sixty topics taught in schools. The inter-boy and inter-girl correlations were calculated and factorized. The results were suggestive of two main types, the artistic or humanistic, and the scientific. Type I chiefly liked topics connected with English, French, drawing, etc.; Type II also liked drawing, but put topics connected with physics high and with languages low. Shakespeare (1936) has criticized these results, doubting whether they reveal as much as his own analysis of children's preferences according to age and sex. But it is clear that both approaches yield useful results; there is no need to regard them as rivals.

Stephenson has conducted further experiments on preferences for colours, odours, aesthetic objects, etc. Investigations at the

Department of Psychology, University College, on attitudes to pictures, also indicate the existence of distinctive types of appreciation. Some of this work will be presented in a forthcoming article by Dewar (1938).

(iii) *Conformity—Atypicality of Opinions*

71. A large variety of tests have been based on the extent to which a testee's responses approximate to or deviate from some predetermined standard responses. Thus Deutsch (1923) measured conventionality or conformity by giving a list of questions with multiple choice answers, one of which was the conventional answer (e.g., one question dealt with methods of disposal of the dead). The proportion of conventional answers checked by the testees was found to correlate well with assessments by acquaintances of the testees' conventionality. Artistic or musical taste is almost always measured by presenting reproductions of contrasted works of art (photographs of furniture, selections from poems, gramophone records of music, etc.); these works have previously been judged by art experts, and the number of times a testee's preferences agrees with the experts' preferences is taken as an index of his taste. The validity of these tests is somewhat dubious for reasons which the writer has discussed elsewhere (1935). Barry (1931) attempted to test "compliance" or "suggestibility"; he first gave an opinionaire dealing with a variety of topics to a group of students, and later repeated it telling the students, before they marked each item, what was the previous commonest response in the group. He then summed the numbers of items to which they altered their first responses in the direction of conformity with the group response. Individual differences in susceptibility to propaganda have been investigated similarly.

72. Ethical discrimination tests.—Innumerable tests of moral judgment or ethical discrimination have been devized, the earliest being Fernald's (1912). A good description of them may be found in Symonds (1931, pp. 268-285). Here the testees register their opinions on moral issues and are scored according to the conformity of their responses to some arbitrary ethical standard. Any of the techniques already described may be adopted:—voting on the blameworthiness of various crimes; ranking the ten commandments in order of importance; true false tests, e.g.—

Good marks are chiefly a matter of luck — True — False
Clean speech is a sign of being a goody-goody — True — False

multiple choice items, e.g.—

If someone steals your lunch, you should:
- Steal another lunch to even it up
- Report it to the teacher
- Cry about it
- Say nothing about it.

Similarly, moral stories, or pictures of good and bad deeds, may be presented together with several alternative outcomes, the best of which is to be chosen.

73. Reliability and validity of ethical discrimination tests.—The reliability of these tests is generally high, so long as they contain a large enough number of items, but there is grave doubt as to what they measure, i.e. their validity. Hartshorne and May (1928), who devized and tried out a very ingenious series of them, found that different moral judgment tests gave positive though rather low correlations with one another ; but that there was scarcely any agreement with tests of honest or dishonest behaviour. Others have demonstrated that the moral judgments of delinquent children tend to be just as " correct " as those of non-delinquents. Similarly, when 200 prisoners and 272 school teachers arranged 45 criminal acts in order of seriousness, the two lists were practically identical, showing that deviations from the teachers' list cannot be taken as an index of criminality (cf. Simpson, 1934b). Many investigators find a significant correlation between moral judgment scores and intelligence tests, which suggests that the responses of delinquents are determined more by their comprehension of society's moral conventions than by their own moral habits of behaviour. But why should there be so much less relationship between attitudes and behaviour in this instance than in most of the tests we have described in earlier sections? We would suggest, though we cannot yet prove, that the difference is mainly due to the different *Einstellung* among the testees. They realize, or at least the more intelligent ones realize, what the tester is trying to measure, and assume that it may be to their advantage to give conventionally moral responses. Thus the results of these tests supply a valuable commentary on the dangers of deducing conduct from verbal attitudes, and suggest the extreme importance of the attitude with which the testees approach the test situation.

(iv) *Extreme versus Moderate Opinions*

74. Psychological conception of extremeness v. moderateness.—F. Allport and Hartman (1925), Vetter (1930) and G. Allport (1937) are doubtful as to the legitimacy of the conception of radicalism-conservatism, and point out that in many ways the extreme radical and the extreme reactionary are more like one another than either is like the liberal or moderate person. Similarly, both British and American public speakers and editorial writers are constantly drawing attention to the resemblance between Communism and Fascism, and their opposition to democratic ideals. It should be possible to test this quality of extremeness *v.* moderateness by measuring the *dispersion* of responses in a multiple choice opinionnaire. We have already seen that there are wide individual differences in dispersion of ratings (§ **4**) ; similarly in a radicalism-conservatism test, where five or more grades of response are provided, some persons may check more extreme radical *and* conservative responses, and yet get the same radicalism scores as others who only give intermediate responses. Watson (cf. § **31**) has used this extremeness as a measure of prejudice in his Fairmindedness test. And Thouless (1935)

(cf. § **17**) studied individual differences in certainty (i.e. extremeness) of religious and other beliefs, hoping thereby to obtain a measure of irrational thinking.

75. Experimental justification.—Unfortunately, experimental justification for this conception is as poor as the justification for the radical-conservative continuum is good. Testees do vary widely in extremeness on many tests, but fail to do so consistently. The present writer has often found scores among students ranging from 10 or 20 per cent. to 80 or 90 per cent., where 0 per cent. represents the greatest possible moderateness, 100 per cent. greatest extremeness. Yet the correlations between these scores, derived from different attitude tests, were generally negligible. Moreover, there is little proof that these variations correspond to any distinctive psychological trait, with the exception of sex—men usually giving more extreme opinions than women. Watson's measures do agree fairly closely with his other tests of prejudice ; and in one study the writer found low positive correlations with tests of impulsiveness-caution. Allport and Hartman (1925) claimed to have found personality differences between extremes and moderates, but their results were probably not very reliable statistically. In another experiment the writer compared four different measures of extremeness with his testees' results on the Boyd *Personality Questionnaire* (§ **140**), but failed to find any consistent personality correlates. We must conclude then that though this conception is psychologically promising, much more rigorous experimental investigation of its reliability and validity is needed.

(v) *Variability of Opinions*

76. Harper (1927), Lentz (1930, 1934), and Telford (1934) have studied individual differences in the numbers of altered responses when the same opinionaire is given a second time after a three or four weeks' interval. With a long test these change scores are reliable, but, just as with extremeness, the scores derived from different tests are rather inconsistent. Though there appears to be no correspondence between variability and general instability of personality, there is a slight negative relation with intelligence (the more intelligent being more consistent), and, according to Harper, with liberal opinions.

IV.—ASSESSMENT OF HUMAN TRAITS BY RATINGS

A. Introduction

77. In everyday life we are continually making judgments of one another's character and temperament traits. Not merely in ordinary social intercourse, but also in education, industry and the professions, we realize that these traits are fully as important as intellectual or other abilities. But our judgments are often haphazard and biased, based on quite inadequate knowledge of the people we judge. The various rating methods have developed in an attempt to render them more objective, more systematic, and therefore

more useful both for practical and for research purposes. No accurate and easily applicable tests are available for the assessment of most personality traits, so that we are forced to rely very largely on ratings. And ratings actually possess a considerable advantage over such personality tests as have been devized, in that they can often be applied without the knowledge of the ratees (i.e. the persons rated), whereas the person who knows that he is being tested can hardly be expected to exhibit his normal emotional characteristics. It is probable then that, next to intelligence tests, ratings are more widely used and have been subjected to more thorough study than any other psychometric technique.

78. Uses of ratings.—In many American business concerns, colleges and schools, it has become a matter of routine to check up on the personality characteristics of the personnel, students and children by periodic ratings. Though less developed in Britain, we may note their habitual employment in the National Institute of Industrial Psychology's vocational guidance service, and their adoption in the important researches of Galton on imagery, Pearson on intelligence, Burt on delinquency, and of Webb, Cattell, Stephenson and others on character. The ordinary school report about a child's conduct ("Excellent," "Fair," etc.) is an ubiquitous instance of a bad rating scale. Hamley and his collaborators (1937) have attempted recently to work out a more logical and comprehensive record card for schools, which includes systematic personality ratings. Among the many other spheres where rating methods have proved useful are the assessment of the goodness or badness of factory buildings, and of school textbooks, and in the attitude surveys and tests already described.

79. Comparison between ratings and attitude tests.—Ratings of personality traits and responses to an attitude test are closely similar, not only in the technical methods they employ, but also in the psychological processes they involve. Rater A's judgments of B's personality are expressions of his attitude towards B, in just the same way as his judgments of items in an opinionaire express his radical or other attitude. Generally speaking, we use the ratings as a source of information about B (the ratee) rather than about A (the rater), whereas an attitude test is applied mainly for the information it gives us about A. Nevertheless, as we shall see below, it is important to keep in mind this dependence of the assessments on A's affective judgments as well as on B's characteristics.

B. Rating Techniques

80. Ranking and paired comparisons.—Ranking or paired comparisons may be used when the number of ratees is small, in just the same manner as was described in §§ **7–10.** For instance, a school teacher may be asked to rank the children in her class in order of industriousness, or some other trait. Very careful definition of the trait, according to the principles outlined below (§ **89**), is essential if these techniques are to be effective.

When a number of individuals are to be rated on several traits, and when none of the available raters is acquainted with the whole group of individuals, an alternative procedure is recommended by Burt (1937b), namely, that the raters should rank the order of prominence of all the traits in each individual, instead of ranking or rating all the individuals on each trait. He claims to have obtained more reliable ratings by the former than by the latter method (cf. Burt, et al. (1926). "Most people find it far easier to think of a person's character as consisting in a pattern of tendencies differing in relative strength rather than in a sum of isolated traits to be assessed on a normal curve." The present writer has also found this method useful, but it does not seem to have been adopted elsewhere.

81. Voting Techniques.—The voting technique (cf. § **2**) is exemplified by "Check List" and "Guess Who" ratings. In the former a long list of traits is provided and the rater checks those which appear to him to fit the testee. This is too crude a method unless there are very large numbers of raters; it has, however, been extended and applied in May and Hartshorne's (1930), Thurstone's (1934), and other investigations. May and Hartshorne's (1930) "Guess Who" technique is especially suitable for use with school children. A series of short character sketches is drawn up, e.g. word pictures of a very selfish, a moderately selfish, an average and an unselfish child. These are given to all the members of the class, who are told to fit them to any pupils they seem to describe, i.e. to guess whom they represent. Owing to the large number of raters, a pupil's score for selfishness is readily obtained from the number of times each sketch is assigned to him; and the scores show high reliability.

82. Numerical ratings.—Much more frequently adopted are techniques which regard a trait as a quantitative variable, and assign to each ratee a certain low or high score on this variable. Giving marks out of 20 to one's friends for Beauty, Humour, etc. used to be a popular parlour game. We know now, however, that the average rater cannot properly discriminate twenty grades of a trait. Judgments in terms of percentages have disappeared for the same reason. As already described (§ **3**), the best results are obtained when the trait is divided into 5 or 7 steps or grades. A smaller number than this may be justified (4, 3 or 2) when an analytic rating scale (cf. § **90**) is used; when the raters are numerous, or when they are insufficiently interested in the rating or insufficiently trained to make finer discriminations.

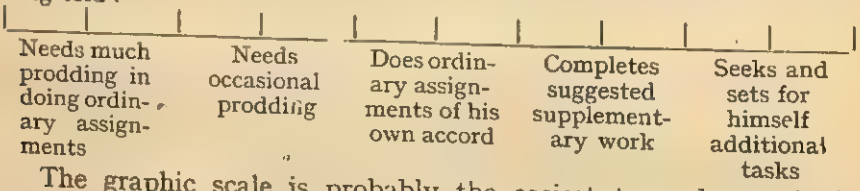
83. Errors in numerical ratings.—We have already dealt (§§ **3–5**) with the distribution of ratings, with the correction of errors in their average and their dispersion, and with the methods of combining the judgments of several independent raters. Most raters are with difficulty persuaded to use the extreme steps in their judgments of people; some are more cautious than others—hence the desirability of specifying the proportions of ratees to be assigned to each step. The error in average level is also very prominent owing to an inveterate tendency to leniency amongst almost all raters. Unless specially

trained, they put far too may ratees above the average on any desirable trait, too few below, apparently regarding an average or 0 rating as something discreditable. On account of these various difficulties which raters find in using a numerical scale consistently, two alternative techniques have been devized—the Man to Man scale and the graphic scale—the second of which is now almost universally employed.

84. Man to Man scales.—This type of scale was developed by Scott (1923) in 1917 for use among American army officers. In rating, say, leadership, each rater was told to think of the officer A whom he regarded as highest in this trait, then another E who was lowest, another C half-way between, and others, B, D, halfway between A–C and C–E. These five names were retained as a private yardstick. In rating any officer X, the rater had to judge which of the five X most closely resembled in leadership. Similar yardsticks were constructed by each rater for several different traits. The method is concrete and provides the rater with a fairly permanent set of standards. But it is somewhat cumbrous, and has the disadvantage that each rater's standards are different.

85. Graphic scales.—This type, first suggested by Freyd (1923), attempts to establish the same standards for all the raters, by careful definition of every step on the scale. The following is an example from a scale to be used by college tutors in rating their students.

Does X need constant prodding, or does he go ahead with his work without being told ?



The graphic scale is probably the easiest to understand, the quickest to apply, and the most interesting for untrained raters. The rater now no longer needs to think quantitatively; he simply writes a check mark at a position on the line which represents X's standing. But the experimenter can measure off the position of this mark in as accurate quantitative terms as he wishes. It is found that five inches is the best length for the line, since it obviates crowding of the descriptions, and yet can readily be grasped as a whole. By an appropriate choice of descriptions, the "central tendency" (tendency not to use extremes), and the tendency to leniency can be partially controlled (cf. Symonds (1931), Guilford (1936)). E.g., a more symmetrical distribution of ratings will be obtained from :—

than from :—

86. Two further precautions may assist in securing more careful judgments from the raters. They should rate all the ratees on one trait at a time, not rate each ratee in turn on all the traits. And the desirable and undesirable ends of the scales should be placed at random on the left- and right-hand sides of the page in successive traits. Uhrbrock (1932) even recommends "scrambling" the descriptions of the various steps in a graphic scale, for example :—

Very good
Fair
Excellent
Poor
Good

so as to force the raters to read all the descriptions before they decide on their rating. This plan is however not generally adopted except in the following, Thurstone-type, scales.

87. Rating scales in equivalent units.—As with attitude tests, the numerical ratings obtained by the above techniques are not scaled in equivalent units. We usually assume that the distances between successive steps are equal, though we have no justification for so doing. Given sufficient raters and ratees it is possible to establish a rational scale by Thurstone's methods (**§§ 11–13, 45–48**). For instance Richardson and Kuder (1933) constructed a scale of 51 statements for rating the efficiency of Proctor and Gamble's salesmen, a scale which they claim to be highly reliable. The following are sample statements, with values on a scale of 0 to 8 units.

(6·9) He is making exceptional progress
(3·2) He is somewhat in a rut on some of his brand talks
(5·6) He tends to keep comfortably ahead of his work schedule

Any of the statements that are thought to apply to the ratee are checked, and their average scale value gives his efficiency score, in scientific units. Willoughby's Emotional Maturity scale (cf. § **137**) is similarly standardized. Most investigators, however, seem to regard such refinements as unnecessary.

88. Social maturity scales.—Scaling according to age level is used by Bridges (1931) in her rating scales for emotional and social development of pre-school children, and by Doll in his recent Vineland Social Maturity Scale (1936 ab, 1937). (It should be pointed out that psychological age units are *not* equivalent). The Vineland scale contains 117 items, such as :—

Reaches for familiar persons
Dries own hands
Is trusted with money
Makes telephone calls
Provides for the future

It has been proved that the first item represents a level of maturity attained by normal (American) 4 month old babies ; the second by 2½ year old children ; and the third, fourth and fifth by 5½, 10½ and 25 years respectively. While the scale can be used for any investigation involving assessments of social competence, it was first devized for the purpose of grading feeble-minded persons who, as is well known, often differ more from normal persons in social than in intellectual maturity. Though the Stanford-Binet test is invaluable for measuring the mental age (M.A.) of defectives, it needs to be supplemented by assessments of their emotional and social develop-

ment. The scale is applied by a trained examiner, who obtains the necessary data from a competent informant in an interview; he checks the items in the scale that apply to the patient and then works out the Social Age (S.A.) and S.Q. in the same way as a Stanford-Binet M.A. and I.Q. Excellent reliability is claimed for the scale (cf. § **234**).

C. Traits to be Rated

89. Avoidance of ambiguous traits.—In the early days of rating it was common to find very general traits such as *Leadership*, *Efficiency*, *Honesty*, etc., being assessed. The defect of these is that they require so much interpretation on the part of the rater; different raters may attach a variety of different meanings to them. Hence such scales tend to show poor reliability, or lack of agreement between raters. Numerical or letter ratings, or adjectives such as "very high, superior, poor," etc., are also liable to produce confusion and variation of standards. American textbooks always insist therefore that the traits to be rated and the steps on the scale should be as specific and objective as possible. Each set of ratings should deal with a single type of behaviour which has been actually observed by the raters and which can be assessed reliably. Adequate definition of a trait or type of behaviour is not achieved merely by listing several synonyms, but rather by describing in easily understood terms concrete situations in which the trait is expressed to various degrees. And impersonal situations (where the ratee's trait in some way affects the objective environment) are preferable to personal ones (where he only creates an impression on his social environment). According to Hollingworth (1929), some traits are rated worse than others. Character and moral qualities such as *Courage*, *Unselfishness* and *Integrity* are far too equivocal; whereas *Perseverance*, *Efficiency*, *Quickness* and the like are relatively definite and can be rated more reliably. It is generally believed, however, that better results with all traits are obtained by the use of analytic scales.

90. Analytic scales.—Here, instead of asking for ratings of general efficiency, we split it into the particular actions that go to make up efficiency, and have these rated separately. The rater's notion of the ratee's general trait was presumably derived originally from a multitude of observations of items of behaviour; and by using analytic scales his general impression is dissected so as to get at the facts on which it was based. The analytic scale is the analogue of the opinionaire test of an attitude, with its series of separate items which represent concrete manifestations of the general attitude. As in the opinionaire then, the items of an analytic rating scale are eventually recombined to give a total mark for the general trait; and they may be weighted according to their relative importance as components of the trait. Again, while they should of course be statistically consistent with this total mark, they should be relatively independent of one another, so as to sample as widely as possible the whole field covered by the trait (cf. §s **36e, 39, 56**).

91. Limitations of Behavioristic approach to ratings.—The emphasis on Behaviorist principles in ratings, which is reflected in the last two paragraphs, is perhaps exaggerated among American psychologists. Certainly it is desirable to aim at concrete descriptions of the steps on our scales, but we cannot hope to eliminate all interpretation. There is, indeed, no definite dividing line between subjective interpretation and objective observation; the rater's recollections of the ratee's actions will always be coloured by his general emotional impressions, and his notions of the ratee's general traits will always be mixed up with memories of particular actions. In the following discussion we will attempt to give the evidence for this standpoint, and to show its bearings on the practical problems of rating scale construction.

First it should be pointed out that Behaviorist definitions are very seldom achieved. The graphic scale for "Working without prodding," quoted in § **85**, is indeed fairly objective; but many scales include highly interpretative items. The following, from Laird's (1925) *Personal Inventory C 3* for introversion-extraversion is an example :—

Has X (in the past few months)					
experienced fine sentiments and emotions ?	easily moved to tears	sympath-ised readily	not especially sensitive	often unmoved	unsympa-thetic

92. Rating of observed or inferred behaviour.—Let us consider the claim that ratings will be useless unless the raters have had opportunities of observing and collecting relevant evidence about the type of behaviour to be rated. For instance, a teacher should not be asked to rate a child's *Physical Agility*, nor his parents to rate his *Concentration of Attention*. This principle appears to be sound, and yet it neglects the tendency of human beings to express many facets of their personalities in everything they do, and the tendency of judges to interpret the person as a whole rather than merely to observe his specific behaviour. Newcomb (1931) has demonstrated that ratings on observed items of behaviour are but little superior to ratings on behaviour which has only been inferred. In his investigation at a boys' summer camp, careful records were kept of the actual behaviour, and the accuracy of subsequent ratings was checked against these records. The accuracy of ratings on observed items was represented by correlations of $+0{\cdot}54 \pm{\cdot}09$ and $+0{\cdot}45 \pm{\cdot}10$, on unobserved items $+0{\cdot}39 \pm{\cdot}11$ and $+0{\cdot}40 \pm{\cdot}10$, figures which are very little lower.

93. Reliability of different types of ratings.—The criterion most frequently applied for determining the goodness of ratings is the reliability (consistency), i.e. the extent of correlation between different raters' judgments of the same traits. Actually this criterion does not always show conclusive superiority for analytic scales over ratings of single general traits. In several studies reported in the literature the consistency of analytic scales was represented by

correlations of about + 0·50 or less; though some more recent scales have yielded better results, namely, coefficients of + 0·70 or more. In early researches by Burt (1915), ratings were obtained on MacDougall's list of primary emotions; also systematic records were compiled of specific manifestations of these emotions. The consistency of the former averaged about + 0·45 to + 0·50, of the latter + 0·60 to + 0·70. Newcomb's investigation, mentioned above, showed much the same consistency among ratings of observed and of inferred behaviour. However, the significance of this criterion is far from clear, since a high correlation may merely show that the raters happen to have common prejudices about the ratees (cf. § **106**), not that their judgments are true ones.

94. Wolf and Murray's investigation of ratings.—Some relevant evidence was obtained in a recent investigation by Wolf and Murray (1936). They made an extremely thorough study of the personalities of 28 students. Each student was first rated on forty traits by five judges after a 45 minutes' interview, then tested or analyzed by various investigators for a total of 36 hours, and finally re-rated on the same traits in the light of all the information which had accumulated. Taking these final ratings as criteria, the initial ratings showed the fairly high average validity of + 0·63. But it was found that those traits on which the initial raters most closely coincided were not necessarily the most validly rated, though there was some positive relation between consistency and validity. Next, it was noted that ratings on "latent" traits, i.e. the deeper personality trends which require most interpretation, were but little inferior in consistency or validity to ratings on "manifest" traits, which are more directly expressed in overt behaviour. On the whole, however, traits which were well defined and understood, and which were best evoked during the interview (e.g. *anxiety*, *emotionality*) were more validly judged than were the more obscure traits, and those which were seldom evoked (e.g. *creativity*, *sex*).

95. Observational and time-sampling techniques.—Before concluding this section we must mention the important observational techniques which have been developed in America by Olson, Thomas, Goodenough and others, chiefly for studies of pre-school children. Instead of drawing on interpretative judgments or recollections of behaviour, accurate records of strictly defined types of behaviour are made at the time of their occurrence. It is seldom possible to do this continuously; hence what is called time-sampling is adopted. Olson and Cunningham (1934), who give a useful summary of applications of time-sampling to some forty types of behaviour, define it as: "systematic recording of a definitely delimited unit of behavior described in terms of action over a stated time interval, yielding quantitative individual scores by means of repeated time units." Thus, in studying social participation at a Minnesota nursery school, Parten (1932) observed 34 children for one minute each a day, for sixty days, distributing each child's minutes systematically throughout an hour's school period. In the minute she checked whether the

child's behaviour consisted of solitary play, onlooker behaviour, organized group play, or other defined categories of sociality. A child's score was derived from the total number of minutes (out of 60) in which he had been engaged in each category. This and other similar studies generally discover remarkably high reliability; a child's social participation score on odd-numbered days correlates closely with his score on even-numbered days. Olson (1929) has applied the method to older children at school, recording their nervous habits such as nail-biting or tics, and their whispering, during specified time samples. Thomas, Loomis and Arrington (1932) have carried out extensive observations in a garment factory, recording the activities of an operative at 5 second intervals continuously for 5 minutes. By confining themselves, however, to a Behaviorist classification of activities ("Material, Self, Contact with Persons, Non-job Activities, and Language"), they seem to have obtained results of purely technical, not psychological or practical, interest. Nevertheless, time-sampling might be profitably extended to industrial investigations of work concentration, talkativeness, posture, etc*. In addition to its accuracy, it has the merit of measuring spontaneous and natural behaviour in the kindergarten, school or factory environment, without any of the artificial restrictions that are so often unavoidable in psychological tests or laboratory experiments. The testees need not even know that records are being taken if suitable one-way observation screens are used.

96. Reliability (consistency) of observational records.—Our immediate concern with these techniques lies in their consistency. In general the correlation between two observers who make simultaneous records is decidedly higher than the consistency of ordinary ratings. Moreover, Thomas claims that observations of highly specific objective activities (e.g. total number of contacts with other children) are more consistent than those of activities which involve interpretation (e.g. number of social contacts). But such specific fragments of behaviour have very little relevance for personality until they are interpreted. Total contacts may be very accurately measured, but they are psychologically meaningless. And Goodenough (1928, 1930) shows that social participation, leadership and the like *can* be accurately recorded if they are sufficiently clearly defined, and if the recorders are sufficiently trained.

97. Conclusion.—These various researches, together with others described below, appear to lead to the conclusion that ratings will be somewhat improved by careful and concrete definitions of the traits or types of behaviour, and that it is probably worth while substituting for each general trait a *short* series of relatively objective component scales. But unless we confine ourselves to techniques such as Thomas's, we cannot hope to get strictly factual ratings,

* Such techniques are, in effect, already employed by industrial psychologists in studying output and accidents. The discovery of Farmer and others that records of accidents sustained during a sample period of a year are highly predictive of subsequent accident rates, is a good instance.

freed from all subjective interpretation. Thus a very detailed Behavioristic dissection of a trait into an analytic scale of many items is unnecessary, and does not confer any advantage to compensate for the great increase in labour which it demands from the raters.

D. Influence on Ratings of Extent of Acquaintanceship

98. Ratings based on photographs or the voice.—Many studies have been made of the accuracy of ratings when the amount of information concerning the ratees is severely limited. For instance it is important to know how much can be deduced from facial expression alone, as portrayed in a photograph, how much from a brief interview, and so on. The standard method is to obtain ratings on several traits from a group of judges who have interviewed, or looked at photographs of, the ratees, and then to correlate their judgments with ratings on the same traits obtained from close acquaintances. Alternatively intelligence may be rated and the judgments compared with intelligence test results, or judgments of vocation compared with actual vocation. Similar judgments of traits, vocations, etc., may also be based on hearing the voices of the ratees. The results with photographs or the voice alone are uniformly poor (cf. Pintner (1918); Landis and Phelps (1928); Hollingworth (1929); Taylor (1934), and a dozen other studies). Judgments of intelligence or of character traits tend to give zero correlations, unless the ratees are a specially selected group. Burt (1919) shows however that emotional and physical traits, such as *Humour* and *Muscular Energy* are slightly better rated.

99. Difficulties in these experiments.—Doubtless the main reason for these negative findings is that a single photograph provides so limited a vehicle for the expression of personality; it may or may not portray the characteristic facial pose. Uhrbrock (1930), and the present writer, have found that judges are much influenced by the chance factor of whether the ratees happen to be smiling or frowning. Most raters tend to assess the former more highly than the latter in intelligence or in any other desirable trait. (Candidates for jobs who are told to send photographs with their application blanks are therefore advized to make sure that they are photographed smiling). A further difficulty is that the judges of the photographs, and the acquaintances who provide the criterion ratings, may not attach the same meanings to the traits that they rate. Another factor, which the writer has discussed elsewhere (1936), is that most judges do not readily think of personality in terms of a series of separate quantitative traits. The poor validity of their judgments may then be due to the artificial form into which they are forced to compress their deductions from the photographs or the voice. Burt's method of ranking traits within the individual ratee (§ **80**) and May's "Guess Who" technique (§ **81**) are two attempts to substitute more natural forms of judgment, which appear to possess

advantages over the ordinary rating techniques. A third possibility is known as the matching technique.

100. The matching technique.—Here, for example, the judges are given brief sketches of the personalities of the individual ratees, and instructed to identify or match these with the appropriate photographs or voices. Several continental psychologists—Binet, Arnheim, and Wolff (whose work is summarized in the writer's article, 1936)—obtained high proportions of successful matchings by this technique. The writer has devized a statistical method for expressing the validity of such results in terms of contingency coefficients, the application of which indicates that judgments of personalities considered as integrated wholes are indeed superior to judgments of separate traits in a group of persons. Thus in Allport and Cantril's (1934) experiments on the voice, judgments of separate traits or interests yielded an average validity contingency of $+0{\cdot}27$; but when the same lists of traits were combined into brief character sketches, the matching of these with the same voices gave a contingency of $+0{\cdot}41$; (the probable errors of these coefficients were approximately $0{\cdot}01$ to $0{\cdot}02$, so that the superiority is certainly significant). The matching technique is most suitable for the study of somewhat tenuous indices of personality, such as photographs, the voice, gestures, and artistic style. Its extension to judgments of more complex material has been but little explored.

101. Ratings of dynamic appearance : interviews.—The living and moving features naturally offer much more scope for judgments than do photographs. Cleeton and Knight (1924) found low positive correlations between judgments of students sitting, silent, on a platform and criterion ratings supplied by their acquaintances. Estes (1937) obtained similar results from ratings of motion pictures, and these were somewhat improved when the method of matching with character sketches was adopted.

In interviews, many additional deductions may be made from the ratee's answers to questions; hence we usually, though not invariably, find better validity. Wolf and Murray's result has already been quoted (§ **94**). Burt (1919) obtained coefficients averaging $+0{\cdot}33$ between ratings of children after brief interviews and their teachers' ratings on the same traits. This figure should be contrasted with the average of $+0{\cdot}12$ obtained from ratings of photographs. Magson (1926) found scarcely any agreement between judgments from interviews and tests or ratings of intelligence, but ascribed this partly to the diversity of interpretations of the meaning of the term intelligence. Hollingworth (1929) has supplied a devastating demonstration of the inability of experienced business men to agree upon the applicant who would be most suitable for a particular job. Twelve experienced judges interviewed each of 57 candidates and arrived at entirely discordant conclusions. Webb (1915), on the other hand, found as high agreement between ratings by interviewers and ratings by teachers who knew the ratees well, as between the interviewers among themselves or the teachers among themselves.

102. Ratings of temperament and personality on the basis of behaviour during testing and interviewing.—Still more data are available when the ratees are given something to *do* during the interview. If they are set some baffling task, their persistence, reactions to success and failure, etc., may be very revealing. This is the situation in the individual application of the Binet-Simon, or performance, tests of intelligence. Burt (1926), Stutsman (1931), Anderson (1929), the present writer (1929) and others have described the many qualities that may be observed under these conditions. The writer has prepared a graphic rating sheet, listing these qualities, which he has found useful in interviews and individual tests with adolescents and adults ; it is reproduced on pp 56–57.

Burt and the writer find good consistency between different observers who judge the same testees under these conditions ; and the writer has obtained an average validity coefficient of $+0{\cdot}50$ by comparing observers' ratings on several traits with a series of independent measures of these traits. Though this figure is only moderately high it compares well with the validity of ordinary ratings by associates (cf. § **110**). A much higher figure can hardly be expected, since the testing situation is necessarily different from, and more restricted than, situations of everyday life. Many testees may be upset by its novelty and fail, for instance, to persevere at the tasks set them, although usually very persistent at things which interest them elsewhere.

103. Ratings in vocational guidance interviews.—The vocational guidance technique worked out at the National Institute of Industrial Psychology combines such observations of test behaviour with an informal discussion or interview about the candidate's previous history, present interests and future ambitions (cf. Rodger (1934)). This interview is specially designed to elicit information bearing on the personality traits which will play an important part in vocational adjustment, and is therefore likely to be of much greater value than the haphazard interviews which led to Hollingworth's poor result (cf. § **101**). Thus it seems likely that the observations plus the discussion should supply sufficient data for judgments of quite high validity, although, of course, the knowledge of the candidate's personality that is obtained in the limited time available must be far from complete. Excellent evidence as to the validity of the ratings (in terms of which the examiner sums up the candidate's traits) is provided by the successful outcome of this type of guidance (cf. Myers, 1932). Another factor which probably helps to improve the judgments is the experience that the examiner has gained in maintaining an impartial attitude towards the candidates. The iteration of his job, and the very briefness of the mutual acquaintanceship, may be advantageous in preventing the growth of subjective prejudices for or against the candidates which, as we shall see below, so greatly distort ordinary ratings by close associates.

104. Ratings by acquaintances.—Ordinary ratings are not usually based on any immediate observations, but rather on the raters' recollections of all their past experience of the ratees. Often they will have seen the ratees in much more varied situations and so have a much wider range of data to draw on than in any of the investigations we have been describing. There is still of course some restriction; for acquaintanceship is generally confined to one phase of people's lives. For instance, the foreman has little opportunity to observe the factory worker's home life, and the relatives do not know how he behaves at the factory. We have, however, already pointed out the weakness of the view that raters can rate accurately what they have actually observed, and nothing else (§ **92f.**). Subjective generalization and interpretation probably play a large part even in estimates of objective characteristics of the ratee's behaviour. We tend to see people and to note their actions through the distorting spectacles of our own affective reactions to them. In an interesting study by Landis (1925) of the reasons which raters give for their judgments, it was found that the ordinary rating is not based on definite past observations, but rather on a general impression of a diversity of experiences. The reasons were usually very vague, often bizarre; they might well be termed rationalizations, in the psychoanalytic sense.

105. Conclusions as to the effects of acquaintanceship.—We need no longer be surprised that prolonged observation and close acquaintanceship do not necessarily improve ratings. The disagreements between different acquaintances rating the same individuals are very striking. According to Symonds (1931) the typical correlation between the judgments of a pair or raters is about $+0{\cdot}55$. Though higher figures are sometimes obtained, they may often be lower if the raters are untrained, the scale badly prepared, and the traits obscure. The preceding Sections do indeed indicate that an increase in the amount of data concerning the ratees generally increases their accuracy up to a point. But the results of Landis (1925) and Shen (1925) show that intimacy of friendship has no effect, and Knight (1923) found evidence of poorer judgments of ratees who had been longest known to the raters. Apparently a relatively superficial and detached knowledge may be superior because prolonged familiarity tends to make the raters stereotyped and biased in their attitudes. It is sometimes recommended that raters should carefully study the ratees for a period before they make their judgments, keeping in mind the traits to be rated. (This plan was adopted by Webb (1915)). Though certainly desirable in so far as it can be carried out with impartiality, yet we must be prepared for the resulting increase in knowledge to be offset by the development in the raters of greater prejudice.

General Rating Scale for Qualitative Observations during Testing and Interviewing

Name .....Date ..... Examiner .....

Activity

Excited, restless, unable to keep still	Impulsive
Quick and vivacious	
	Stable
Calm and deliberate	Cautious
Inert and listless	Inhibited

Poses, motor attitudes .....

Tics .....Nail-biting.....Twitchings.....

Fiddling with

material ..... clothes .....hands..... feet.....

Peculiar expressions.....

Excessive wrinklings.....

Movement

Fluent and graceful	
Accurate and well-controlled	Quick stride and movements
Angular and awkward	Slow stride and movements
Clumsy	

Physique and Bearing

Impressive in bearing	Healthy looking, well developed and nourished
Satisfactory impression	
Unimpressive	Unhealthy, feeble physique

Forceful, efficient, energetic, upright posture and gait	
Slouching gait	
Weak, inefficient movements and bearing	

Plump (pyknic) proportions	Florid
Well and symmetrically proportioned	
Thin (asthenic)	Pale

Personal Appearance and Expression

Attractive and good-looking (positive reaction)	
Pleasant	
Uninteresting, indifferent attractiveness	Sensual .....
Ugly and repulsive (negative reaction)	Effeminate .....

Strong expressiveness of face and gestures	Frank
Expressionless	Secretive
Quick and strong sense of humour	Cheerful, optimistic
Slow but sure	
Unable to see humour	Depressed, melancholy
Mature, serious, philosophical	Excitable, irritable
Immature, childish	Even-tempered
	Calm, phlegmatic

Special Characteristics .....

.....

.....

SPEECH

Voice resonant, pleasing, well-modulated	Clear, fluent, distinct
Hard, harsh, pinched	Stutters, stammers

Expresses meaning directly, grammatically, with facility
Unable to express himself, ungrammatical

Accent.....

Garrulous, over-talkative	Brilliant in talking, wide vocabulary
Rather voluble	Dull and stolid, narrow vocabulary

Seldom speaks of own accord
Reticent, taciturn

PERSONAL CARE

Fastidious in dress, over-manicured
Good taste, neat and clean
Passable and inconspicuous
Careless in dress and cleanliness
Slovenly and unkempt

SELF-ASSERTION

Pompous and overbearing	
Complacent	Decisive
Self-confident and possessed	
	Wavering
Self-critical and deprecatory	
Embarrassed, bashful, self-conscious	Contrasuggestible
Anxious, apprehensive	
Submissive, retiring	Suggestible

CO-OPERATIVENESS

Willing to co-operate in every respect ; enters into spirit

Reserved and formal
Constrained and suspicious, outside the situation
Surly and hostile

Scrupulous, punctual and regular in attendance, application
Industrious
Easy-going, indifferent
Lazy and irregular

ALERTNESS AND CONCENTRATION

Intelligently attentive, wide-awake
Concentrated

Absent-minded
Easily distracted, inattentive

TEST REACTIONS : PLANNING

Analytical	
Serious but unsystematic	
Trial and error	Profits by past experience
Haphazard	
	Repeats same mistakes

EMOTION

Wild and unrestrained emotional behaviour and remarks
Wilful and childish reactions, capricious
Some loss of self-control, and overt emotion
Humorous and unconcerned
Serious, philosophical
Repressed and inhibited

E. Halo Effect and the Reliability and Validity of Ratings

106. Halo effect.—Early in the evolution of rating methods it was found that sets of ratings on different traits when inter-correlated almost always yield unduly high coefficients. A pair of traits which appear to have no *a priori* connection with one another are yet discovered to possess some prominent common factor. For instance, in an investigation by the writer, ratings of students by their friends on *Sociability* and *Quickness of Movement* gave a correlation of + 0·71. Thorndike (1920) suggested that the common factor is the general impression of, or attitude towards, the ratee possessed by the raters, which colours all their judgments of particular traits. For instance, we regard A as highly artistic, and tend to attribute to him all the other traits commonly associated with the artistic temperament; or we think of B as generally melancholy, and fail to notice such of his characteristics as do not fit in with our over-simplified notion. Most commonly, it would seem, halo consists largely of our general liking for, or our dislike of the ratees. For it is usually found that the desirable or admirable traits give high positive inter-correlations, and negative correlations with undesirable traits. Doubtless this has some basis in actual fact; persons of fine character do tend to be high on all good qualities, others do tend to be weak all round. But we are very liable to exaggerate this, and to attribute unwittingly all the virtues to our friends, all the vices to our enemies. Hence we fail to distinguish properly between traits which should be relatively discrete. Uhrbrock (1932) has shown that raters who are personally selected by the ratees themselves are especially apt to over-rate on all desirable qualities.

Many intelligent and conscientious raters no doubt realize the existence of this effect and try to allow for it, to dissociate their personal prejudices from their judgments. Yet even when forewarned it seems to be impossible for raters to avoid it altogether. The effect is most pronounced when poorly defined general traits are to be rated, especially traits that bear ethical connotations; but it is present also even in the rating of objective items of an analytic scale, and produces some spurious correlation between the items. According to Knight (1923) it increases with intimacy of acquaintanceship, since this makes us all the more blind to defects in our friends or to good points in our enemies.

107. Influence of halo on reliability and validity.—Undoubtedly then we should be extremely cautious in accepting ratings as true measures of the traits of the ratees. We are unable to cite here all the evidence; but Argelander (1937) has recently provided a good account of: "the dependence of judgments of human character on the personality of the judge in his functions as observer, interpreter and evaluator." It is found for instance that different raters agree more closely with one another when they possess a common slant towards the ratees. A schoolchild tends to possess

much the same personality in the eyes of two of his teachers ; his parents may also look on him alike, but the agreement between parents and teachers may be relatively small. This is partly due of course to the different environments in which teachers and parents see him, and his different behaviour in these environments, but it also reflects his different haloes. May and Hartshorne (1930) show that it is better to regard ratings primarily as the *reputation* of the ratees in the eyes of the particular set of raters. To do this need not imply that ratings tell us nothing of value about the ratees ; indeed we shall see below that they do agree moderately closely with measures of behaviour. But it does mean that we make allowance for the subjective element, just as we are accustomed to do in everyday life. We do not usually take one another's opinions about other people at their face value, but normally interpret them in the light of our knowledge of the judges. A factory worker can tell us something about the personality of his foreman, a mother about her son, though we realize that they are both looking through coloured glasses. Similarly then, ratings should be interpreted with full regard to their origins.

108. Improvement of reliability and validity by pooling ratings.—Although the judgments of a single rater have very little worth as measures of an individual's traits, yet it is clear that the pooling of judgments from different raters will tend to cancel out some of the subjective errors and so lead to ratings of greater reliability and validity. It is generally considered desirable to combine at least five raters ; though the number varies with their experience. It is less often realized that raters should be chosen who have seen the ratee from as many diverse viewpoints as possible. For instance, the combination of ratings on a child from two teachers, two parents, and two of his friends is likely to give a much better picture of him than a pool of six teachers' judgments ; biases which will be reduced by the former plan will tend to reinforce one another in the latter. In other words we should try to get a representative sample of the ratee's various reputations.

109. It should be noted that this scheme will not bring about so high a reliability (consistency) coefficient as will the pooling of the estimates of similar raters, although it will tend to improve their validity. We have here the same phenomenon as with attitude scales (cf. § **56**) ; a very high consistency may indicate uniform bias rather than approximation to the truth. Again, when a second set of ratings is obtained from the same raters, a high repeat reliability is desirable, but it should not be taken as a criterion of accuracy, since it could equally well result from fixity of prejudice.

Reliability, it is generally claimed, may be improved not only by increasing the number of raters, but also by increasing the number of ratings through the use of analytic scales. We have already dealt with this point (§§ **90–97**), showing that some subdivision of traits into concrete items is worth while. We also noted that the different items should be (like the different raters) relatively independent of one another.

110. Empirical evidence of validity.—There have been few studies of the validity of ratings, in the sense of comparisons with other more objective criteria of people's traits. Much more often they are assumed to be valid, and are used as criteria for determining the validity of some test of aptitudes or of character. However, the work of Hartshorne and May (1928), together with unpublished work by the writer, where large numbers of tests as well as ratings were applied to the same subjects, do tend to show that an ordinary set of pooled ratings is superior to any single personality test or short battery of tests. The writer obtained an average validity coefficient of $+0{\cdot}60$, whereas most of the alleged personality tests yielded figures around $+0{\cdot}30$ to $+0{\cdot}45$. It is obvious from the previous discussion that we cannot expect measures of reputation and measures of behaviour to cover precisely the same ground any more than an attitude test can accurately predict a testee's actions.

111. Co-operation and training of raters.—There is a further possible flaw in ratings, namely, that they are not necessarily perfect measures of reputation. For just the same inhibitions, hesitations, misunderstandings, etc., are likely to be aroused in the rater when he is asked to assess an acquaintance on a rating scale, as when he is asked to fill in a political or religious attitude scale (cf. §s 57–58). Obviously, he will seldom give his candid opinions if he has the slightest fear that the ratee will see these opinions. And his willingness to put down what he believes to be the truth may vary considerably with the social respectability of the trait; e.g. he might admit that the ratee was extremely impulsive, but shrink from attributing to him extreme conceitedness or dishonesty. Clearly then it is desirable to take into account the attitude of the raters towards the task and towards the experimenter, and to secure as complete co-operation as possible before the rating begins. Conrad (1932) recommends that the rater should be interested in, and should realize the usefulness of, the ratings; also that he should be allowed to take his own time over them. Kingsbury (1922) supplies useful hints on the training of raters. The scales that are to be employed should be discussed so as to ensure that the definitions are clear. Instruction with regard to the main sources of error, and practice in the application of the scale, should help to reduce the central tendency, the leniency tendency, and the halo effect.

F. Factor Analysis of Ratings

112. The inter-locking of ratings.—As early as 1906 Heymans and Wiersma studied the overlapping between ratings on a large number of traits. For instance, they selected from their ratees those described as persistent liars and found what other traits were ascribed to this group, how emotional, how active, how perseverative they were, and so on. Though their technique was clumsy they foreshadowed what is now one of the most popular lines of research, namely, the discovery of common factors running through sets of traits by factorial techniques. Allport and Odbert (1936) have

recently listed 17,593 personality traits. Naturally many of these are synonymous, and still more are likely to overlap. For instance, there might be a common element in "emotional, unstable, impulsive, excitable, temperamental, variable," and a dozen other such terms. Factorizers hope then to distinguish a relatively small number of discrete dimensions of personality into which these thousands of traits may be resolved.

113. Attempts to remove the halo factor.—Now ratings provide the easiest, and possibly the best single method, of getting measures of traits. But we have seen that all sets of ratings start out with the very influential common factor of halo, which is responsible for a high degree of spurious overlapping. As Guilford (1936) states, until this is removed or corrected, nothing can be deduced from trait inter-correlations. No satisfactory method for removing it has yet been devized, though several possibilities have been suggested.

(*a*) Either the sum of all the ratings on desirable traits, or the first general factor to be extracted in analyzing the correlations should be regarded as measuring halo, and should be held constant by the partial correlation technique. This is scarcely fair because, as pointed out above (§ **106**), the positive overlapping of desirable traits is to some extent a real feature of the personalities of the ratees. Nevertheless, in one experiment by the writer, the validity of ratings (compared with other measures of the same traits) was somewhat improved by this expedient.

(*b*) When factorizing measures of abilities we would expect the first, *g*, factor to account for the major part of the inter-correlations. But with measures of varied personality traits, the first few factors should presumably be more nearly equal in size. If therefore, as in most of the studies described below, the first factor is far larger than the others even after rotation of axes, its excessive size might be made to yield an index of halo. The flaw in both these suggestions is that they assume halo to be itself a unitary factor. Actually it is likely to be much more complex than mere desirability of traits, and so to enter more or less into all the factors extracted, though chiefly into the first.

(*c*) Raters might be asked to estimate their personal liking for or dislike of the ratees, in addition to assessing their personality traits, and these popularity ratings might be partialled out from the trait ratings. For the reason just mentioned, this would certainly not eliminate, though it might reduce, halo effect.

(*d*) May and Hartshorne (1930), finding an average consistency of + 0·92 among teachers' ratings of pupils and among the pupils' ratings of one another, but a correlation of only + 0·48 between teachers and pupils, point out that the excess of the former over the latter figure is due to the halo components in the ratings, and that the common ground between teacher and pupil ratings (represented by the lower coefficient) may correspond to the ratees' actual behaviour. They do not show, however, how the halo and behaviour components can be effectively separated. More recently Chi (1937) has attempted to implement the suggestion. He assumes that halo is an *individual* attitude which is different in each rater, A, B, C . . . etc. Thus the correlation between A's ratings of traits 1 and 2 ($r\ A_1\ A_2$) will be spuriously increased by halo; so will all other correlations of the same type (e.g. $r\ A_1\ A_3$ or $r\ B_1\ B_2$). But the correlation between A's rating of trait 1 and B's rating of trait 2, and others of this type ($r\ A_1\ B_2$, $r\ A_2\ B_1$, etc.) will not be affected. In a detailed investigation he has worked out by this means all the trait inter-correlations freed from halo ($r\ A_1\ B_2$, etc.) and all those due to halo ($r\ A_1\ A_2$—$r\ A_1\ B_2$, etc.). Unfortunately his major premise cannot be accepted; halo is not purely individual; it is only too likely to be common

to several raters, especially when, as in this instance, all the raters were teachers. Either then the method must be extended, in accordance with May and Hartshorne's original suggestion, to *groups* of raters; or else a number of entirely independent raters, who are unlikely to possess any halo in common, must be obtained. We might then arrive at a good measure of the extent of halo and of the inter-trait correlations when it is removed. Until this is done, the interpretation of the results of the following factor analysis studies is very hazardous.

114. Webb's investigation.—The pioneer investigation in this field was carried out by Webb (1915), using ratings of schoolboys and students. By Spearman's technique, he extracted a factor which seemed to represent general strength of character or will, and called it *w*. It was most strongly loaded with traits such as "conscientious, persistent, energetic, tactful, emotionally steady, and kind on principle." Webb could not have been aware at that time of the pervasiveness of halo; but it is clear now that the traits listed are precisely of the desirable kind which his raters would be most likely to attribute to those whom they regarded through a favourable halo. How far then *w* should be considered as a fundamental dimension of the ratees' personalities, and how far it corresponds to their mere popularity, cannot be decided.

115. McDonough's, Cattell's and McCloys' investigations. McDonough (1929) adopted Kelley's extension of Spearman's method in analyzing a series of ratings of young children. She extracted four main factors which appeared to represent *Will, Cheerfulness, Sociability* and *Sthenic Emotionality*. Cattell (1933) studied the overlap between ratings on 48 traits which were taken from descriptions of extrovert-introvert, cyclothyme-schizothyme, and other such personality types. The main factor, which he called *Surgency-Desurgency* was especially prominent in the following traits, "cheerful, natural, sociable, humorous, adaptable, gregarious, quick apprehension, etc."

McCloy (1936) took some of Webb's results and also a set of ratings on 43 traits obtained from his own students, and applied to them Thurstone's technique. After rotation of axes, fairly similar sets of four factors emerged from the two sets of data. The first, like Webb's *w*, consisted of all the socially and ethically desirable traits. The second was most highly loaded with aggressive or dominating traits; the third embodied the sociable and extraverted, as opposed to individualistic and non-cooperative traits; the fourth was highly loaded with "energy, health, and physique," but also incorporated a number of logically incoherent traits.

116. Thurstone's investigation.—Thurstone (1934) chose sixty traits to represent as wide as possible a range of human characteristics, and got 1,300 persons each to rate one acquaintance on them. Five independent factors were found to account for most of the inter-correlations; the main traits that went to make up these factors were:—

I. Self-important, sarcastic, haughty, grasping, cynical, quick-tempered.
II. Friendly, congenial, broadminded, generous, cheerful.
III. Patient, calm, faithful, earnest.
IV. Persevering, hard-working, systematic.
V. Capable, frank, self-reliant, courageous.

Thurstone admits that the first factor consists largely of halo. In this instance it is negative in direction and includes all the undesirable traits which raters ascribe to those whom they dislike.

117. Kelley's, Tyron's and Chi's investigations. Kelley (1934) applied Hotelling's technique to the analysis of tests and ratings of children's "courtesy, fair play, honesty in school work, loyalty to fellows, mastery, poise, regard for property rights, and school drive." The two main components seemed to correspond to general social conformity or good citizenship, and individualism or assertiveness. Tryon (1933) obtained different main factors in the two sexes from ratings by children of one another. The chief constituents of the first factor in boys were "active, humorous, friendly, leader, fights, daring, assured, happy, enthusiastic," and in girls, "popular, happy, enthusiastic."

Chi removed the individual halo element from teachers' ratings of pupils by the method described above (§ **113**), and then found a prominent general factor, similar to Webb's, which he called volition in relation to the school environment. Among the correlations due to halo, he found an independent factor which was difficult to identify; it was most heavily loaded with traits such as "persistence and facing reality," least of all with "muscular co-ordination."

118. Burt's investigations.—Burt (1915, 1938) has studied the inter-relations of assessments of primary emotional traits (e.g. fear, rage, joy, etc.), both among normal students and children and among delinquents and neurotics. He used not only ratings but also records of observed behaviour, which were classified under the headings of the various emotions (§ **93**). The latter, while not of course entirely immune from distortions of subjective judgment, should contain much less halo than the ratings analyzed by other investigators. The full results are not yet published; but apparently they provide clear evidence of a general factor of emotionality running through all the separate traits, and a secondary factor named sthenic *v.* asthenic emotions, which corresponds roughly to the conventional dichotomy of extroversion-introversion. The same factors also emerged when correlations between persons (§ **68**) instead of between trait assessments were analyzed.

119. Conclusion.—This survey of the main factor studies in the field of ratings shows some concordance, but also a good deal of divergence between the results of different experimenters, who start out from diverse lists of traits and use diverse statistical techniques. Several of them may perhaps be reconciled if we take the point of view that they are primarily analyses of raters' opinions rather than of ratees' traits; that is, if we regard them as showing the main ways in which different groups of people tend to think about personality. The primary factors in Kelley's and Chi's studies show which from among the traits presented to the raters are most appreciated or depreciated by teachers. Tryon's results similarly reflect the views of boys and girls, and McCloy's and Thurstone's illuminate the psychological thinking of college students. At the same time there is undoubtedly some overlapping between these factors derived from ordinary ratings, and those yielded by Burt's more objective data; there are resemblances also to the products of the self-rating tests, described in the next chapter (§ 122). So that though we cannot accept the claim that these studies have isolated the fundamental components of personality or character, nor decide at present how far they reveal distinctive types of behaviour (as

contrasted with reputation), yet they do provide an interesting and promising method of attack upon problems of personality.

F. Indirect Measures Derived from Ratings

(i) *Judging Ability*

120. If the judgments of a group of raters are studied, it is always found that some agree much more closely than others with the pooled results. Those whose opinions coincide best with the opinions of the group are commonly said to be the " best judges of personality," or to show the most " insight." But if we bear in mind the halo effect, we see that " insight " might equally well be interpreted as " conformity." For goodness or badness of ratings is measured in just the same way as conformity or atypicality of opinions is measured in an attitude scale (cf. §§ **71–73**). Thus suppose that rater A happens to possess a very complete and unbiased knowledge of ratee X, but B, C, D and E have a uniform, yet biased, attitude to X ; A's ratings would then turn out to be the worst. Sometimes a rater's assessments can be compared with a somewhat more trustworthy criterion than ratings (cf. Vernon (1933a), Wolf and Murray (1936)), but most investigations of judging ability have been based on this rather dubious standard.

121. Consistency and validity of judging ability.—Contrary to expectations there is found to be extremely little consistency in the ability. A judge may rate one trait well, another badly ; or he may rate X accurately and Y inaccurately. Either there is no such thing as *general* intuitive ability, or else it cannot be measured by this approach. However, there is fairly good evidence that in the long run better judges are slightly superior in intelligence, in artistic inclinations, and in introverted, asocial tendencies (cf. Adams (1927); Vernon (1933a); Estes (1937)). The latter finding may indicate that the extraverted, sociable person is less capable of standing back and viewing others impartially. In a study by Hollingworth (1929) raters and ratees were one and the same group. It was then found that those who were rated highly on a desirable trait were themselves better at rating it in others, and that the opposite held with undesirable traits (e.g. the most " snobbish " were bad at rating " snobbishness "). Wolf and Murray (1936) have also shown that judges rate best those who are most like themselves.

122. Factor analysis of judging ability.—Burt points out that when we inter-correlate different raters' judgments of a group of ratees, we are actually correlating " persons " rather than " tests " (cf. § **68**). Stephenson (1936b) also discusses this point, and contributes a factorial study of raters by his Q-technique. He obtained from ten teachers judgments of the *Reliability* of a hundred children, determined the correlations between the ten sets of judgments and factorized them. Two main types of raters were found. Those in the first type agreed closely with one another as to the children who were most and least reliable ; those who were saturated with the

second type agreed well among themselves, but did not correlate with the first type. The first appeared to base their judgments of *Reliability* chiefly on placid, submissive behaviour, whereas the second type looked for more active and direct evidence of the trait. We seem then to have here a tool of considerable value for analyzing rating abilities.

(ii) *Judg-ability*

123. As early as 1908 it was found that some ratees were more "judg-able," i.e. more consistently rated than others. Allport (1937) suggests that those about whom the judges agree most closely are more "open," less "enigmatic" in personality; and he finds a definite correlation between measures of "open-ness," and ratings on traits such as extraversion and expansiveness. Some caution is however needed in interpreting this result, for judg-ability is likely to be fully as unreliable statistically as is judging ability. Moreover, it may be that those about whom the raters disagree are, more simply, the ratees who have no very outstanding traits, who are "mediocre" rather than "enigmatic." For it has frequently been noted that extreme ratings are more consistent than intermediate ones (cf. Cady (1923); Hollingworth (1929), etc.), also that those judgments about which the rater feels most sure (and these are usually very high or very low) are superior to judgments given without assurance. All these researches, however, need to be repeated with scales in which the units have been made equivalent by psychophysical techniques (cf. § **87**). Lack of assurance about ratings near the middle of the scale may be merely a function of the units, which are not generally based on discriminable differences.

(iii) *Empirically Standardized Scales*

124. The following technique assumes great importance in some of the tests to be described below, though it has as yet been applied to ratings in only one instance. Its main feature is the introduction of some external criterion for determining what the scale or test measures, instead of relying merely on the apparent meaning of the scale. Clearly the application of such an empirical check on the scale's validity is highly desirable; unfortunately, there is often a tendency, as we shall see later, to eschew all psychological considerations of the significance of the scale, and to trust wholly to the empirical criterion to give results.

125. Olson's scale.—Olson (1930) wished to devize a rating scale for personality maladjustment, or behaviour disorders, to be used by teachers; but was anxious also to avoid giving them a scale which dealt directly with such disorders, since it might be very liable to distortion by halo and other prejudices. Instead, he constructed an ordinary scale of ratings on 35 fairly innocuous and common traits. This is published as the Haggerty-Olson-Wickman *Behavior Rating Schedule B* (1930). Its empirical standardization was carried out by applying it first to a special group of children (the "standardization group"), who had already been rated very

thoroughly on a special analytic scale for personality maladjustment by raters who could be trusted. Taking next all those children who had received some particular rating on the general scale, he calculated their average score for personality maladjustment from the special scale, and so developed *a maladjustment index for that rating*. Similar indices were determined for each of the other possible ratings in the general scale.

Now when a teacher or group of teachers rates any child on the general scale, the child's personality maladjustment may be found from the sum of the indices pertaining to the various ratings he receives. As Olson points out, the same technique could be employed for estimating children's probable scholastic achievement ; the same general scale would be used, but each rating on it would be assigned a scholastic achievement index on the basis of its previous application to children of known high or low achievement.

126. Discussion of the empirical technique.—The technique may seem unnecessarily roundabout, or even perverse, since it takes no account whatever of the conventional meanings of the ratings on the general scale. The significance of such ratings as measures of personality maladjustment or scholastic achievement is determined wholly empirically. We will not try to assess its value until we have seen what results it gives with other types of test ; but would point out here two important requirements if it is to be made to work. First, the initial measure of maladjustment or achievement must be as accurate as possible. In Olson's investigation the initial measure was derived from ratings on the special scale, and is therefore of uncertain validity. Secondly, the standardization group must be very large, for otherwise it is found that the standardized test yields hopelessly inconsistent results. The consistency of Olson's final test was fair ; for when tried out on a new group it gave a correlation with ratings on the special scale of + 0·62.

It is worth incidental mention that the extreme ratings on Olson's general scale almost always obtained the highest maladjustment indices. For instance, on the trait *Physical Output of Energy*, those ratees who are checked as "extremely sluggish" score 5 for maladjustment ; those called " over-active, hyperkinetic " score 4 ; the intermediate steps, " slow in action, moves with required speed, and energetic–vivacious " receive indices of 3, 2 and 1. Child Guidance experts would probably agree that most extreme traits may be indicative of poor adjustment. In this instance then the empirical method has given quite meaningful results.

V.—SELF-RATINGS AND PERSONALITY QUESTIONNAIRE TESTS

A. Description of Tests

(i) *Introduction*

127. Self-rating tests, personality " inventories " or " schedules," or " pencil and paper tests " (as they are sometimes called) follow precisely the same lines as individual attitude tests and analytic

rating scales. Their object is to obtain in quantitative terms an individual's estimates of his own character and temperamental traits. In a number of experiments, individuals have been instructed to rate themselves on the same scales with which they rate others. It is commonly found that, while an individual's associates may differ considerably among themselves in rating him, his judgments of himself tend to be still more divergent from the group judgments; and that he is especially apt to over-rate himself on desirable traits, i.e. to possess a favourable halo towards his own character. Since then self-ratings, either on a single trait, or on a series of traits, are of such dubious significance, an analytic scale technique is much more frequently adopted.

128. These analytic scales include a large number of items (anywhere from 10 to 223) bearing on the general trait, and they are therefore much more likely than single self-ratings to yield reliable measures. Moreover, the items cover a wide range of presumed manifestations of the trait; they can deal primarily with questions of the testee's behaviour rather than with his self-evaluation, and they may be to some extent disguised (in much the same way as in the *Study of Values* or Watson's *Fairmindedness Test* (§§ **30–31**)). Hence their validity should also be improved.

It is probable that a hundred or more of such tests have been published. But the great majority are simply modifications or extensions of three prototypes: Woodworth's *Personal Data Sheet*, Freyd-Heidbreder's *Introversion-Extraversion Test*, and Allport's *Ascendance-Submission Test*. We will, therefore, outline the origin and construction of these three, and mention a few others which have been widely used, or which embody special points of technique.

The form of test items or questions is the same as in attitude tests or graphic ratings. All of them have been standardized by one or other of the three methods we have already described: the internal consistency technique (§§ **35–43**), empirical techniques (§§ **124–126**), or the external judgments and Thurstone-scaling technique (§§ **44–53**). Almost all of them are intended to be used as group tests.

(ii) *Tests of Emotional Instability or Psychoneurotic Tendency*

129. Woodworth's Personal Data Sheet or Psychoneurotic Inventory.—The 116 items in Woodworth's test were originally derived from descriptions by psychopathologists of the symptoms of neurotic patients (e.g. from J. T. MacCurdy's *War Neuroses*). The following are some representative examples:—

Do you usually feel well and strong?
Do you ever walk in your sleep?
Have you ever had fits of dizziness?
Did you have a happy childhood?
Do you know of anybody who is trying to do you harm?
Does it make you uneasy to cross a bridge over a river?
Have you ever been afraid of going insane?
Has any of your family had a drug habit?

Each question is followed by "Yes No", one of which is to be checked.

130. Mathews's, Laird's and Thurstone's tests.—Mathews (1923), Cady (1923) and others have adapted the test for use with children. Laird's (1925) *Personal Inventory B2* contains a similar series of items, but with multiple choice (graphic) responses, e.g. :

Have you (during the past few months) been afraid of responsibility?	avoided it	accepted when forced upon me	did not mind it	liked it	welcomed it

Thurstone's *Personality Schedule* (cf. Thurstone and Thurstone (1930)), which is now the most widely used, contains 223 items collected from Woodworth, Laird, and other sources.

131. Standardization and scoring.—In these, and the many other derived tests, it has usually been shown that the items do hang together consistently, by the internal consistency technique. Or items may be selected from a more extensive preliminary draft on the basis of their differentiation between testees who obtain extreme scores on this draft. Thus Willoughby's (1932b) *Clark-Thurstone Schedule* includes the 25 best differentiating items from Thurstone's *Schedule* ; graded responses (0 to 4) are provided.

The testee's score usually consists of the total number of items (unweighted) to which he responds in the psychoneurotic direction ; it is a " width " rather than an " altitude " type of score (cf. § **48**). Percentile norms are generally provided for showing whether his score should be deemed high, low, average, etc.

132. Burt's Questionnaire on Neurotic Symptoms.—Burt (1937c) has published an English adaptation of the Woodworth and Thurstone tests, but does not intend it to be scored quantitatively. He employs it simply as one device for eliciting qualitative information about personality problems in an interview.

(iii) *Tests of Introversion-Extraversion*

133. Freyd-Heidbreder's test.—Freyd (1924a) collected 54 items descriptive of the introvert type from Jung's writings, of which the following are samples :—

Blushes frequently ; is self-conscious.
Daydreams.
Prefers to read a thing rather than experience it.
Shrinks when facing a crisis.
Is reticent and retiring ; does not talk spontaneously.
Is slow in movement.
Keeps in the background on social occasions.

Heidbreder (1926) turned them into a self-rating test ; the testee is instructed to check each item +, ? or —, according as it applies to him or not. Laird's (1925) *Personal Inventory C2*, and a host of other adaptations are available.

134. Tests standardized on psychotic patients.—Some psychologists, distrusting the internal consistency technique, have attempted to apply an external criterion of the trait in selecting their items, i.e. to adopt an empirical technique. Neymann and Kohlstedt

(1929) worked on the rather doubtful assumption that schizophrenic or dementia praecox patients represent the extreme of introversion and that manic-depressive patients are extreme extraverts. Their test contains the fifty items which were found best to differentiate between two such groups of patients. A shortened form, prepared by Root, was used in Wyatt's (1937) research, and is reproduced in full in Report No. 77. The Neymann-Kohlstedt, and other tests which have been similarly standardized, tend to give very poor correlations with introversion-extraversion tests that have been standardized on normal persons by the internal consistency technique; conversely, the commoner form of test differentiates poorly between such mental patients.

(iv) *Tests of Ascendance-Submission and Other Personality Traits*

135. The Allport A–S Test.—In this test (cf. Allport (1928)), the items were devized as concrete manifestations of dominatingness (ascendance) or submissiveness, e.g. :—

A salesman takes manifest trouble to show you a quantity of merchandise; you are not entirely suited; do you find it difficult to say "No"?	Yes, as a rule..... Sometimes ..... No .....
If you hold an opinion the reverse of that which a lecturer has expressed in class, do you usually volunteer your opinion?	In class..... After class..... Not at all.....

An alternative form is available for women, and an adaptation for children has been prepared. Empirical standardization was used. Allport first obtained ratings of students by their associates on ascendance-submission; he then applied a preliminary draft of the test to those who were rated as most ascendant or most submissive, and, on the basis of their responses, chose the best items and calculated an appropriate weighted mark for each response. The marks for the above items are — 1, 0, + 1, and + 3, — 1, — 3, respectively.

136. Tests of other traits.—A test for *Inferiority Feelings* based on Adler's writings, has been compiled by Heidbreder (1927), along the same lines as her *Introversion-Extraversion Test.* Bernreuter (1933c) has published a test of *Self-sufficiency v. Dependence on Others,* Jasper (1930) a test of *Depression-Elation,* which are similar in form to those already described. Cason (1930) gives a list of 217 *Annoyances,* i.e. situations which tend to annoy people. If a testee rates each of these from + 3 (extremely annoying) to 0 (not annoying), his average rating can be used as a test of *Irritability* or *Annoyability.*

(v) *Tests Based on External Judgments*

137. In Wang's (1932a) test of *Persistence,* 111 items were selected from a larger number in accordance with the views of 75 judges, who decided whether each item described a persistent or non-persistent person. The Thurstone scaling technique was adopted by Willoughby (1932a) in constructing his *E–M* (emotional

maturity) *Scale.* He collected 150 items descriptive of symptoms of emotional maturity or immaturity, e.g. :—

S develops affective difficulty in the presence of a necessity for precise or realistic thinking, e.g. mathematics _____

S organizes and orders his efforts in pursuing his objectives, evidently regarding systematic method as a means of achieving them _____

These were sorted by 101 professional psychiatrists or psychologists, according to their degree of maturity. Sixty items were retained ; the two reproduced here possess scale values of 2 and 7 on a 1 to 9 scale of equivalent units. This test is intended primarily for third-person application, e.g. ratings of a patient by a psychiatrist, but has also been used in the first person for self-ratings.

Chant and Myers (1936) present a similarly constructed scale for *Depression-Elation* ; its items range from :—

Everything in the world is against me (Scale value 0·9) to : Life could not be better for me (10·7).

The authors show that this gives good differentiation between manic patients, normals and depressives.

Both these scales yield " altitude " scores, based on the average scale value of the endorsed items.

(vi) *Multiple Tests*

138. Classification of psychoneurotic-test items.—Tests such as Woodworth's or Thurstone's obviously contain a wide range of symptoms drawn from many distinct neurotic or psychotic conditions. It would be possible for several testees to give psychoneurotic answers to entirely different sets of, say, 20 items. Although this would indicate that they were completely different from one another, yet they would all get the same score and so be labelled equally unstable or psychoneurotic. Surely it would be more logical to group the symptoms into separate sets, each set centred round one main type of personality disorder. Harvey (1932), Willoughby (1932b) and others have attempted to classify the Thurstone items, the latter using as his categories Social-Asocial, Extravert-Introvert, Fantasy, Physical Disorders, Parental and Sex. However, on intercorrelating scores on these six sets, he obtained little evidence that the sets were really distinctive (cf. § **145**). In Laird's *B2 Inventory* the items are classified as either Psychasthenoid, Schizophrenoid, or Neurasthenoid.

139. Cattell's tests.—Cattell (1936) has published a questionnaire test with separate sets of items for seven pathological syndromes—Neurasthenia, Anxiety Neurosis, Anxiety Hysteria, Conversion Hysteria, Obsessive-Compulsive, Epileptoid and Paranoid. Realizing the difficulty (which we discuss below) of obtaining candid responses to such intimate questions, he intends it only for intelligent and highly-co-operative testees, or else for use in the third person by a clinician rating a patient. In another test, called a *Projection*

Test, he eliminates all direct questioning of the testee, as shown by the following item :—

John strained every nerve to beat the others because :
he was determined to be top.....
his father wished him to succeed.....
he needed the scholarship.....

The testee is told to check the most appropriate of the three endings. It is assumed that he will tend to project his own chief impulses into the endings he chooses. Thus, if he is very self-assertive, he would be likely to check the first in this example ; if submissive he would prefer the second. The test contains 74 such items which are classified so as to yield scores on Self-Assertive *v.* Submissive tendency, Cautious *v.* Bold, Acquisitive, Gregarious, Curiosity, and Dependent or Appeal tendency. This classification is adopted from MacDougall's list of instincts. The theoretical basis of the test is ingenious, but it is unstandardized, and the practical results are so far dissappointing.

140. Boyd's Personality Questionnaire.—This test (unpublished) is apparently the only one that has been widely used as a group test in Britain. Its 120 items are classified under twenty headings or general tendencies, including the following :—

Tendency	*Sample Question*
Obsessional carefulness	Do you often go over a job again and again to make it just right ?
Worry, Anxiety	Do you brood long over humiliating or unhappy experiences ?
Suspiciousness	Do you sometimes suspect that people are talking about you ?
Self-consciousness (Introspectiveness)	Are you greatly interested in what goes on in your own mind ?

The testee is, however, not told about these tendencies, and the questions are so arranged that he is unlikely to guess that six of them deal with carefulness, six with worry, and so on. Questions may be answered Yes, Yes ?, 0, No?, or No, or omitted. These are marked 4, 3, 2, 1, 0 and 2 respectively ; so that the scores on each tendency may range from 0 to 24. Whether the twenty tendencies are really distinctive and self-consistent will be considered below (§ **146**).

141. Maller's Character Sketches.—Maller (1932) has adapted the Guess-Who rating technique (cf. § **81**) in a self-rating test for adolescents and students. Two hundred short descriptions are given, and the testee has to say whether or not he feels or acts like the person described. This impersonal form is apparently much less disturbing than direct questions. Examples are :—

_____ This person insists on having his own way and likes to command and rule everybody
_____ This person finds it difficult to forget unpleasant memories and can't help thinking about them.

Every item is repeated elsewhere in the test, but in reverse form, e.g. :

_____ This person never insists on having his own way and does not like to command and rule everybody

By this means the piling up of descriptions of " unpleasant " people is avoided (cf. §**158**) ; also the carefulness of the testee can be checked by noting whether he answers each reversed question in the opposite way to the original. In the process of standardization it was proved that all the questions differentiated significantly between groups of 310 normal pupils and 308 problem cases, delinquents, etc. The questions are classified under six headings :

Desirable character traits
Self-control and integration
Social adjustment (extraversion)
Personal adjustment (freedom from anxiety)
Mental health (freedom from psychotic or neurotic symptoms)
Readiness to confide in others.

The extent to which these different categories overlap is not stated ; but their average intercorrelation is $+ 0{\cdot}38 \pm {\cdot}08$.

142. The Humm-Wadsworth Temperament Scale.—This test (1935) is perhaps the most logically worked out of any we have described. It aims to measure the seven main components of temperament distinguished by Rosanoff : Normal, Hysteroid or Anti-social, Manic Cycloid, Depressive Cycloid, Autistic Schizoid, Paranoid, and Epileptoid. After preliminary trials, 318 items were chosen and tried out on groups of patients, criminals and normals who were known, from psychiatric diagnoses, to be strong or weak on these components. Marks for the items were calculated from their results. Nearly half the items failed to differentiate significantly between any of these groups ; but they were left in the final form of the test for the sound reason that the differentiating items might be altered by their omission. The testee's response to an item is often determined not by that item alone, but also by its context (cf. footnote § **35**). This procedure has the disadvantage of making the scale rather long ; the average time needed for answering is 55 minutes.

It is claimed that the scores obtained on these seven tendencies differentiate well between further groups of mental patients. The extent of overlapping among the scores of normal persons is not stated. By noting the proportion of Yes and No responses to the test as a whole, a good check is provided on the conscientiousness of the testee. Negativistic persons tend to give an undue proportion of No's ; highly suggestible persons give too many Yes's.

143. The Bernreuter Personality Inventory.—This test (1931, 1933a) was also empirically standardized. But the items were not logically selected in the first place to measure different traits (as they were in Allport's, Neymann-Kohlstedt's, Maller's and Humm-Wadsworth's tests). Instead, they were taken more or less at random from various tests. Each of the 125 items, it is claimed, provides an indication of all the four traits at which the test is aimed, namely, Neurotic Tendency, Introversion-Extraversion, Dominance-Submissiveness, and Self-Sufficiency. The method of standardization was to obtain responses from a group of students who had already taken Thurstone's *Schedule*, Laird's *C2 Inventory*, Allport's *A–S* and Bernreuter's own test of *Self-Sufficiency*. Then for each

of the 375 responses, four marks were calculated on the basis of the agreement of that response with the four previous test results. Thus the answers to the item, "Do you day-dream frequently?" are marked :—

Answer	Neurotic Tendency score.	Intro-version score.	Domin-ance score.	Self-Sufficiency score.
Yes	+ 5	+ 3	− 1	+ 1
No	− 4	− 4	+ 1	− 1
Doubtful	− 2	0	+ 2	− 2

A testee's four scores are the sum of such + and − marks. It will be seen that the construction and standardization of the test are analogous to the procedure used in Olson's rating scale for personality maladjustment (§s **124–126**). Nominally they are wholly objective ; but in point of fact the criteria employed were far from objective, since they consisted of four self-rating tests. Nevertheless, the test works sufficiently well to have achieved tremendous popularity ; it is said that some 50,000 copies of it are sold annually.

B. Multiple Factor Analysis of Self-Rating Tests

144. Overlapping of traits tested by personality questionnaires.— The attempt to classify test items or symptoms logically into distinct groups has not, we must admit, been successful. On the one hand, it is found that tests of presumably different traits intercorrelate very highly; on the other hand, different tests of nominally the same trait, although fairly reliable in themselves (cf. § **168**), tend to give very poor correlations with one another. It is doubtful then whether most of the traits at which the tests have been directed are unitary and discrete. Obviously, the psychoneurotic questionnaires include a hotch-potch of symptoms ; and introversion-extraversion tests appear to cover almost as diverse a collection, e.g. lack of social interests, inhibition of emotional expression, etc. The two conceptions indeed overlap very largely. The present writer has collected from the literature the results of 40 experiments, which show that the average correlation between different introversion tests, and the average correlation between introversion and psychoneurotic tendency tests, are practically identical, namely + 0·36 ± ·10. A further 18 experiments with the *A–S* test show an average correlation of + 0·30 between submissiveness and introversion or psychoneurotic tendency. Tests of inferiority feelings also agree quite closely with tests of introversion. When these traits are combined in a multiple test such as Bernreuter's the overlap is much higher. The average of four experiments where the Bernreuter scores were intercorrelated is* :—

Neurotic tendency with Introversion	+ 0·93
Neurotic tendency with Dominance	− 0·81
Introversion with Dominance	− 0·67

* The *Self-Sufficiency* scores are relatively distinct ; they correlate − 0·41, − 0·32 and + 0·58 with the other three. But the main reason for this is obvious when one examines Bernreuter's *Self-Sufficiency Test*, namely, that it is itself an amalgam of Ascendance + Introversion items. Further comment on these extremely high Bernreuter correlations appears below (§170).

In the Boyd *Personality Questionnaire*, the average intercorrelation of scores for nineteen tendencies, obtained in the writer's experiments, was $0 \cdot 366 \pm \cdot 060$†.

145. Results obtained by multiple factor analysis.—Multiple factor analysis has therefore been applied in an attempt to show objectively which types of items do hang together consistently, and which are independent. In two researches new tests have been evolved to measure the factors that have emerged.

In Willoughby's (1932b) study of the Thurstone *Schedule*, nearly 50 per cent. of the correlations between his six categories of items were accounted for by one factor (which incidentally satisfied the Spearman tetrad-difference criterion). Perry (1934) applied the Bernreuter, Laird *B2* and *C2*, *A–S* and other tests to a group of students and obtained first a large factor which was chiefly made up of Bernreuter Neurotic and Introversion scores and the Laird scores. A second factor seemed to represent a separate Dominance tendency, being highly weighted with *A–S* and with Bernreuter Dominance and Self-Sufficiency.

Flanagan (1935) applied Hotelling's technique to the Bernreuter scores, and found that they were in effect measuring only two distinct tendencies, not four. The first, a compound of Neurotic, Introverted, Submissive and low Self-Sufficiency scores, seemed to him to represent general "Lack of Self-Confidence." This accounted for 78 per cent. of the variance. The second, which covered 18 per cent., he identified as a "Sociability" factor; the third and fourth factors could be neglected, since they were responsible for only 4 per cent. of the variance. Flanagan has constructed fresh tables of marks for each response to the Bernreuter items, so that the test can be scored for these two factors.

146. Multiple factor analysis of the Boyd Questionnaire.—The present writer took the results of 50 men and 50 women students on the Boyd *Personality Questionnaire*, and applied Thurstone's technique to the correlations between the 19 sets of scores. A general factor was found to account for 41 per cent. of the variance (relative to the reliabilities of the scores), and three further factors accounted together for another 35 per cent. The average size of the fourth factor residual correlations was only $\pm 0 \cdot 048$, and none of them was greater than three times its probable error; hence further factorization was not worth while. It follows then that the 19 measures are very far from distinct; each one, moreover, gives appreciable correlations with at least two of the factors.

After three rotations of axes it appeared possible to identify the factors tentatively. The first was prominent in almost all the measures, but especially in the following:—

Depression or Melancholy; Instability or Temperamentalness; Worry or Anxiety; Lack of Self-Control; Shrinking of Responsibility; Lack of Self-Sufficiency or Confidence.

† The twentieth measure was too unreliable to be of any use. This figure is corrected for attenuation.

These all gave loadings of + 0·73 to + 0·57, and suggest a general self-depreciatory or psychoneurotic tendency.

The second factor, independent of the first, might be called a tendency to "Care-freeness," its highest loadings (0·48 to 0·30) being with :—

Shrinking Responsibility; Suggestibility; lack of or freedom from Worries, from Self-Consciousness and Emotional Thinking; Dissociation or Unintegrated Thinking; Inability to Concentrate; Lack of Definite Interests; and freedom from Tenseness.

The third factor, again independent, gave the clearest picture, which might be named "Scrupulousness." With twelve measures the loadings were less than ± 0·10; but with the following seven measures they were 0·60 to 0·24 :—

Obsessional Carefulness; freedom from Instability-Temperamentalness; Acting Readily without Pressure; freedom from Emotional Thinking and from Inability to Concentrate; Suspiciousness; and strong Self-Control of Feelings.

The fourth factor consisted chiefly of those measures in which the women scored higher than the men, or vice versa. It is not easy to interpret, but certainly represents sex differences on this test. The highest loadings (0·56 to 0·22) were, for women :—

Strong dislikes; Strong Fears; Instability; Lack of Self-Sufficiency (i.e. Dependency);

and for men :—

Persecutory (or expecting consideration from others); Suspiciousness; Inability to concentrate; and Self-Consciousness or Introspectiveness.

In judging the significance of these results it should be remembered that the testees were student teachers passing through the strains of adjustment to their career; also that the test certainly does not presume to cover every side of personality, for instance it omits almost all social and moral qualities, interests and values. The factors naturally depend on the types of questions used, and on the people who answer them. Within these limits they do appear to represent rather general and meaningful tendencies.

147. Analysis of introversion-extraversion.—Perhaps the most illuminating research in this field is Guilford's analysis of introversion-extraversion (Guilford and Guilford (1934, 1936)). Taking 36 typical items from introversion tests, he calculated the inter-item correlations and found a fairly prominent general factor, which was chiefly loaded with items indicating sociability or the reverse. Later, he applied Thurstone's technique, and after rotation of axes, obtained five distinct factors, which could be fairly readily identified as :—

I. Social-asocial tendency
II. Emotional immaturity or dependency
III. Masculinity or Aggressiveness
IV. Care-freeness
V. Intellectual interests

(The common conception of extraversion would then be made up of I, III, IV and of the reverse of II and V). Guilford has now published

the *Nebraska Personality Inventory* of 100 items, which can be scored to measure the first three of these factors.

148. Layman's factorial study.—In a detailed research, the full results of which are not yet published, Layman (1937) obtained twelve independent factors from the correlations between 67 personality test items. These items were selected from several tests, and were applied to 276 students. After rotation of axes the factors were identified as follows :—

I. Sociability factor, (i) Gregariousness
II. " " (ii) Feeling of Social Inadequacy
III. " " (iii) Social initiative
IV. " " (iv) Social aggressiveness
V. Changeability of interests
VI. Independence or self-sufficiency
VII. Feeling of inferiority or lack of self-confidence
VIII. Impulsiveness
IX. Emotionality factor (i) Moodiness
X. " " (ii) Sensitivity or excitability
XI. " " (iii) Emotional introversion
XII. Inability to face reality

149. Conclusion.—If we compare these studies we certainly do not find that they yield identical factors ; this could hardly be expected in view of the diversity of the test material with which they started. The majority however agree in showing the predominating importance of one factor, whatever the test or the items used*, which may be interpreted as lack of self-confidence or instability or maladjustment of personality. Several also show an independent sociability factor. When both men and women are tested, there is a sex difference factor. And Guilford and the writer seem to concur rather closely on a " care-freeness " factor†. Layman's study suggests that these rather general tendencies may be split up into several different components when a more detailed analysis is made. It is possible that her Factors I and II correspond to the sociability, III and IV to the aggressive-masculine, V and VIII to the care-free, and the remainder to the general adjustment factors. Thus, although other, somewhat different, patterns of factors are likely to emerge from further researches, yet the results so far obtained do seem to have effected a considerable clarification of the field. We see also that self-rating tests are not likely ever to

* One further study deserves mention, though it did not employ factor analysis. Rundquist and Sletto (1936) constructed six attitude scales for measuring Morale (optimism or pessimism about world conditions and people in general) ; Feeling of Confidence *v.* Inferiority ; Favourableness to the Value of the Family, of Law, and of Education ; and Economic Conservatism. These gave an average positive inter-correlation of + 0·31, and were shown all to measure a general factor which the authors identified with good social adjustment. Apparently then, these tests lead to the same result as the personality questionnaires.

† The writer is indebted to Professor Spearman for the suggestion that this factor corresponds to his *f* or fluency factor. Tests of speed of mental association always correlate highly with ratings or tests of extraversion-introversion ; this may explain why.

be able to cover all the main variables in personality, since items or sets of items which attempt to measure a number of different traits actually indicate only a limited number of distinct dimensions. We shall return later (§ **171**) to the discussion of the psychological significance of such dimensions.

150. Factor studies of annoyances.—Two factorizations of the Cason *Annoyances Test* are somewhat more discordant. Carter, Conrad and Jones (1935) found a large factor of " general annoyability," and minor factors relating to annoyances caused by untidiness, by characteristics of people, etc. Harsh (1936) however arrived at five main dimensions, namely annoyances due to :—

I. Appearance of others
II. Violating of morals or mores
III. Suggestions of superiority in others
IV. Unintentionally disagreeable acts
V. Personal sensitivity.

The lack of agreement may merely reflect the different types of testees employed, and the somewhat different preliminary classifications of the test items.

151. Factorial studies of correlations between persons.—Harsh also applied the " Q-technique " to discover persons possessing different types of annoyability. His full results have, however, not yet been published.

Stephenson has carried out two similar studies with self-ratings. In one of these (1936c) 21 persons rated themselves on 22 traits connected with psychic tempo (Smart, Fluid, Tenacious, Pedantic, etc.) Two independent types of persons sufficed to account for the inter-correlations. The first group seemed to correspond to the normal type, regarding themselves as " lively, fluid, smart, tenacious," etc., and rating themselves low on " inhibited, moody, distraught." The second group was less easy to interpret ; they rated themselves high on " inhibited, comfortable flow of energy, and distraught," low on " fluid, bustling, fanatical, flighty and pushing."

152. In the other research (1936b), ratings were obtained from normals, from manic-depressive and from schizophrenic patients on the relative prominence in themselves of thirty moods (" cheerful, sensitive, nervous, affectionate, serious," etc.). Distinctly different patterns of moods were found in the abnormal groups, though the members of each group correlated closely among themselves. The normals were saturated with these two types to varying extents. Stephenson suggests that a test of cycloid or schizoid tendency could be constructed by correlating such self-ratings with the characteristic patterns of the patients. A test along these lines would indeed possess several advantages over the ordinary self-rating tests described above ; but the proposal has not yet been followed up.

C. Reliability and Validity of Self-Rating Tests

153. Objections to personality self-rating tests.—Many persons on first seeing some of these tests tend to react either with ridicule or disgust. To them it seems obvious that candid answers to such

intimate questions will never be given, so that the results will be entirely worthless. In most countries also psychologists manifest a similar distrust. As mentioned above, Burt (§ **132**) uses his test only as an introduction to a personal interview. In much the same way the vocational guidance examiners at the National Institute of Industrial Psychology obtain from the candidate self-ratings on a list of traits, which are then discussed in the interview (cf. Rodger, 1934). No quantitative results with these, nor with Cattell's or Boyd's personality questionnaires have been published so far. Very occasional reports have appeared of the translation and application of Woodworth's or Thurstone's or other inventories in France, Germany, Spain, Poland, Australia and China.

154. Uses of self-rating tests.—The contrast in America is very striking. Hundreds of investigations have been carried out with such tests, mainly on University students, but also on children, mental hospital patients, delinquents, and other groups. Some Child Guidance and University Clinics employ them regularly, to aid in the diagnosis of maladjustment. In many hundreds of schools in the Middle-West they are applied to all the children of a certain age in the hope of picking out problem cases, with what success we do not yet know. Often also they are used as criteria for assessing the validity of other alleged personality tests, e.g. for investigating the significance of handwriting, or of endocrine records, as measures of personality traits. But the commonest type of study consists of little more than the tabulation of the scores, or of the responses to particular items, given by special groups such as academically successful and unsuccessful students ; criminals ; identical and non-identical twins ; spinsters ; married couples and divorcees. Whether such studies have revealed anything new or important, or whether the tests can validly be applied in clinical practice, is rather doubtful. Indeed it would seem as though American psychologists, flushed with the success of their methods of measuring intelligence and aptitudes, have incautiously assumed that emotional traits could be measured in the same way. A count of the number of intellectual tasks which a testee can accomplish does indeed provide an index of his intelligence ; but a count of the number of symptoms which he checks in the Woodworth or other inventories is not necessarily an adequate measure of his instability or psychoneurotic tendency.

155. Justification of the tests.—Nevertheless we cannot simply dismiss these innumerable investigations as utterly worthless. They are, after all, making use of an extremely valuable source of psychological data, namely the testee's opinions about himself and his recollections of his own experiences, a source which is not tapped either by more objective tests or by associates' ratings. In the present writer's opinion, sufficient results are now available to show that these self-rating tests do measure something, though not perhaps what the authors of the tests generally assume.

156. Importance of testee's attitude.—The outstanding feature of the tests is that so many of their items deal with matters of

strong emotional tone and personal significance, matters which few of us are ready to reveal to the public gaze. We might discuss them with a sympathetic and trusted friend or psychoanalyst, but would naturally hesitate to commit them to writing which some relatively unknown experimenter will read. Thus all the inhibitions, suspicions and difficulties mentioned above (**§s 21–25, 57–58**) are likely to be greatly intensified. Until recently most American psychologists paid little attention to these subjective factors, believing that so long as a standard objective situation was enforced among all the testees, results would be obtained whose validity could be checked by purely empirical methods. But the paucity of such results seems now to have converted many to a realization of the importance of the testees' attitudes to the tests.

157. Conditions affecting testees' attitudes.—Thus Maller (1932) admits that 70 per cent. of students were irritated by the usual type of inventory, and so substituted the impersonal form of question in his *Character Sketches* (**§ 141**), which only irritated 43 per cent. In several experiments he found that a careful preliminary talk on the value of the test and on the desirability of frankness increased the numbers of maladjusted symptoms which the testees admitted; that when they were told that the test would be used to determine vocational fitness, or when they had to sign their names instead of answering anonymously, the maladjustment scores decreased. Olson (1936) and others have also obtained different results from signed and unsigned tests, though the statistical significance of the differences was dubious. Layman (1937) asked her testees which items they would answer differently if they had to sign their names; these items were such that a frank answer "might tend to belittle the individual in the eyes of the social group."

158. Socially acceptable and unacceptable items.—Uehling (1934) criticizes the internal consistency technique of standardization because it leads to a piling up of the most "unpleasant" items, and tends to eliminate the more innocuous. He advocates retaining several questions which may have poor predictive indices but which may calm down the testees. (Such non-diagnostic items are often referred to as "jokers"). The comparison by Smith (1932) and by Rundquist and Sletto (1936) of results obtained with socially acceptable or positive items and unacceptable or negative items (cf. **§ 33**) is very illuminating. Smith included items such as the following in a test of inferiority feelings:—

> Feels people speak well of him and like him.
> Feels people criticize him and dislike him

He found that a considerably greater number would admit that the former (positive) type *did not* apply to them, than the number who would admit that the latter (negative) type *did* apply. The two types inter-correlated very poorly. Rundquist and Sletto find the negative type to be more highly consistent than the positive, and consider that this is because they all arouse a suspicious, hostile or evasive attitude; hence, testees answer them all alike, regardless

of their real meaning. Positive items, on the other hand, are considered more calmly, and so evoke a greater diversity of response and lesser consistency. It is noticeable that the majority of questions in the most commonly used tests of psychoneurotic or introverted tendencies are of the negative type; though some, such as the Neymann-Kohlstedt, *E–M*, *A–S*, and *Character Sketches*, include both types in equal proportions.

Bernreuter (1933b) has shown, however, that the average testee does not merely check the socially desirable responses in his *Inventory*, nor answer simply in accordance with a self-ideal. For in an experiment where instructions were given to respond in these specific ways, the scores were quite different from the scores obtained with the ordinary instructions.

159. Not all the questions, of course, are as intimate or as negative as "Do you feel that life is a great burden?", or, "Have your relations with your mother always been pleasant?". Some of the Woodworth-Thurstone items, e.g. "Do you have a great many bad headaches?" and many of the questions in Allport's *A–S* test deal with comparatively unemotional matters, with facts of past history, or with present physical characteristics. The difference between these two extremes has been demonstrated in several investigations. Neprash (1936), Lentz (1934) and Johnson (1934) analyzed the repeat reliability of items, and found that the more objective ones were most reliable, whereas items involving judgments of the testee's own mental states were most subject to change. Willoughby and Morse (1936) applied a 40-item inventory in individual interviews, and noted the spontaneous comments vouchsafed by the testees. Items "concerned with superficial or conventional matters which touch no affective springs" aroused little or no comment, whereas items which "touch complexes of high affective content—sex, guilt, fear of disapproval . . . " frequently aroused amused, resentful, or embarrassed reactions. There is, of course, no complete separation between these two types: affective reactions and factual statements will intermingle in the same way that interpretation and observation overlap in the rating of others (cf. §§ **91**, **104**). To quote Willoughby again: "There is evidence that a substantial minority . . . will misinterpret or rationalize their response to almost any item, but particularly to those on which an unfavourable or direct response would engender subjective pain." These conclusions are all the more noteworthy in that Willoughby has been responsible for several of the most extensive mass-investigations with personality questionnaires (1934–1936, 1937). Yet he now admits their "susceptibility to complete vitiation by compensatory mechanisms," and the entire lack of control of the attitude of the testees towards the test and the tester.

160. The influence of unconscious affective factors.—We do not merely have to reckon with conscious resistances, hesitations, and lack of candour among the testees. Whether or not we are favourable towards psychoanalytic doctrines in general, we must admit that

psychoanalysts have demonstrated also the importance of unconscious resistances. People literally do not know themselves well enough to answer many of the questions correctly. Their responses are only too likely to be rationalizations or unwitting self-deceptions. In an illuminating discussion of personality questionnaires and psychoanalytic techniques, Alexander (1934) shows that the psychoanalyst employs methods for breaking down resistances and getting at the truth which are the very reverse of those used by the tester. Introspection and critical consideration are of little use to the former ; nor would he ever ask direct questions about personal sentiments and complexes until a suitable state of transference or rapport had been set up. Rather he relies on fantasy, dreams and free association, where conscious control is slackened. But in the application of these tests, conscious criticism is at a maximum. All that the tester can do is to ask for good co-operation, and promise that the results will be treated confidentially (cf. Vernon (1934b)). Again the tests only allow 2, 3, or at most 5 possible responses to each item, whereas the natural reactions of the testees will be infinitely varied. We have learnt also from psychoanalysis that words are a very inadequate medium in which to express our emotional tendencies, presumably because these tendencies are much older and more primitive, phylogenetically and ontogenetically, than are our language capacities. How much more difficult must it be then to express them satisfactorily in terms of Yes, No, or numbers, and the like.

161. Effects of testees' co-operativeness.—We will attempt now to specify the main subjective factors which are likely to influence the testees' responses, factors which are, however, irrelevant to the experimenter's aims. First will come the conscious feelings of irritation, suspicion and resentment, which may lead in some cases to deliberate falsification. These will of course depend largely on the testees' attitude to the experimenter, the extent to which they respect him, and their notions as to his object in applying the test. Obviously some testees will be much more conscientious and co-operative than others. The effects of these conscious attitudes are well illustrated by two investigations among college students.

Hanna (1934) applied the Thurstone *Schedule* to 179 students who came to the College Clinic for psychological or vocational guidance. Presumably they answered it as frankly as possible in the hope that it would help them. Their scores were later compared with independent estimates by clinical psychologists of their degree of maladjustment, and quite good agreement was found. For instance, of those who checked 0–29 psychoneurotic answers 21 per cent. were classified as maladjusted ; among those who gave 30–74 the proportion was 58 per cent.; and among those who scored 75 or over 79 per cent. were maladjusted. Moran (1935) applied a similar test to 189 students, 41 of whom were classified on the basis of other evidence as maladjusted. But here the test was taken along with tests of abilities, etc., at the beginning of the College year, so that many of the testees may have thought that the authorities would be influenced by the desirable or undesirable picture of themselves that the test conveyed. The result was that there was no appreciable difference between the responses of the well and maladjusted students ;

and subsequent investigation revealed many direct contradictions between the responses and their actual symptoms.

162. Effects of testees' self-analyticness.—Secondly, it would seem that some persons are far more introspective than others and more used to verbalizing their emotional experiences to themselves. We would not claim that they actually know themselves better. Indeed medical psychologists often state that such " self-analysts " are more difficult to psychoanalyze than are more naive and unself-conscious individuals ; and some confirmation for this view may be derived from experiments on the good and bad self-rater (§ **178**). While it is unsafe to generalize too far, it is probable that the former are, on the whole, more intelligent and better educated than the latter. If this is true it would explain the remarkable fact, which has emerged again and again from personality questionnaire studies, that University students and members of the professions score much higher in psychoneurotic tendencies than do the relatively uncultured. Frequently they are found to be as unstable as mental hospital patients, i.e. as neurotics and psychotics drawn mainly from lower social classes. (There seems also to be a slight tendency for the better students to be more neurotic and more introverted, though the experimental evidence is far from unanimous on this point ; among children the reverse relationship is more often found.) Now students and professionals may in actual fact be more neurotic and more introverted than other social groups ; but we would suggest that the main explanation of this result is that they are more aware of their emotional lives, and more willing to admit to themselves, and to the experimenter, the possession of the symptoms which the tests describe. Again, the unsophisticated persons who obtain lower scores may be more stable ; but it is also probable that they do not realize their emotional weaknesses. This emotional self-consciousness factor may then bear but little relation to overt behaviour or other signs of psychoneurosis, and yet influence considerably the number of symptoms which are checked.

163. Effects of testees' suggestibility.—Another factor to be taken into account is suggestion. Investigations in legal psychology and experiments on memory have shown how extremely liable to falsification are our recollections of emotionally toned experiences, and how readily suggestive questioning may lead us to accept experiences as our own which never actually occurred. Thus while the majority of testees may be expected, wittingly or unwittingly, to disguise their emotional weaknesses in answering the tests, others of an hysterical or neurasthenic disposition may greatly exaggerate their symptoms. This would correspond to what we termed above, provocation of the responses by the form of the test (§ **25**). While there is as yet no direct evidence of the operation of this factor, the following three experiments appear to show its operation.

164. Hollingworth's investigation of shell-shocked patients. In 1918 Hollingworth (1920) applied the Woodworth test to groups of soldiers in a mental hospital shortly before the declaration of the armistice, and to others shortly afterwards. The average incidence of psychoneurotic symptoms

was roughly half as great in the latter as in the former group, and Hollingworth ascribes this to their different motivation. The former, being in great fear of eventual return to the firing line, consciously or unconsciously ascribed to themselves many symptoms which were rejected when this fear was removed.

165. Watson's investigation of strict upbringing. G. B. Watson (1934) applied a long questionnaire on emotional development to 210 students, and apparently obtained exceptionally good co-operation and frankness. On the basis of the responses to 17 items he separated those subjects who had had very strict parents from those brought up in liberal homes. From the responses of these groups to the other questions he found that the former believed themselves to have had poorer health, to be less well adjusted at school; they disliked their teachers, had fewer friends, more broken engagements; were more prone to day-dreaming and nightmares, and so on. Now it may be that these results constitute a valuable proof of the claims that clinical psychologists have long been making as to the ill-effects of an over-strict home. Other, more objective, evidence might indeed be brought forward to support Watson's conclusions. And yet the results are so "neat" that it is difficult to avoid the suspicion that the more neurotic students, who are generally sorry for themselves, tend both to rationalize away their maladjustment by presuming that they were badly treated in childhood, to remember more of the unpleasant, fewer of the pleasant, experiences of family and school life, and to ascribe to themselves more present difficulties and problems than do the better adjusted. This is at least a possible alternative explanation, which fits in well with the teachings of medical psychologists.

166. Block's investigation of adolescent worries. Thirdly, we would mention a study of sources of conflict between adolescents and their parents by Block (1937). A list of fifty possible sources was checked anonymously by five hundred 12 to 17 year old pupils in an American city. The instructions were to check only those items which were "seriously disturbing" and which made them "very unhappy." Typical results were that 86 per cent. of boys and 71 per cent. of girls checked: "Won't let me use the car"; and that 75 per cent. of boys and 64 per cent. of girls checked "Pesters me about my table manners." Now the *relative* incidence of these and the other sources of conflict in boys and girls of various ages would appear to be highly meaningful but it is very difficult to accept the *absolute* figures. Surely it is likely that large numbers of the testees, seizing the opportunity to pour out their grievances anonymously, accepted a great many items which were suggested by the test, whether or not these were "seriously disturbing." In none of these three investigations need we imply that there was any *conscious* falsification.

167. Effects of mood.—It might be suspected that the testee's temporary mood would have a considerable effect on his responses, that when depressed he would appear more neurotic, introverted, or submissive than when optimistic. An experiment by Johnson (1934) shows, however, that though this does occur, the influence is quite small, and hardly significant statistically.

168. Reliability of self-rating tests.—In general the repeat reliability and the internal consistency (split-half reliability) of the tests are as high as they are among attitude tests. Coefficients of $+0{\cdot}85$ or more are typical. Lower figures are of course obtained when the number of items is small. In Boyd's *Questionnaire* the average coefficient for the sets of six items was $+0{\cdot}583$. Lentz (1934) and others find alterations in some 15–20 per cent. of items on retesting (some items being more liable to variation than others, cf. § **159**), but many of the changes cancel one another out.

169. Significance of high reliability.—It seems probable that, as with attitude tests (§ **56**), these reliabilities are somewhat spurious; they may be due not so much to real constancy in the testees' traits or consistency in all their symptoms, as to the general emotional attitude towards socially unacceptable symptoms, which Rundquist and Sletto discovered (cf. § **158**), or to the stereotyped notions that testees develop concerning the traits at which the tests are aimed and their own standings on these traits. Willoughby and Morse (1936) consider that high consistency arises "not from conspicuous success in measuring a naturally consistent trait, but from consistent warping of responses in the direction of least subjective pain, i.e. towards consonance with an acceptable self-ideal." That the testee's general self-estimate may influence his responses to specific items was indicated in some minor experiments by the writer, where it was found that testees who knew the object of the test obtained quite unduly high consistency. Usually the name of the trait that the test attempts to measure is withheld; some non-committal title like *Personality Inventory*, or *A–S Study*, is given. But this will not prevent the testee from guessing the trait and answering accordingly. The fact that reliability coefficients obtained from tests of children are generally much lower than those quoted for adults (cf. Symonds, 1931) also fits in with our view. Children's conceptions of their own traits would naturally be less stereotyped, and they would less readily recognize the object of the tests. In addition they are of course likely to be more variable and less integrated.

170. Significance of high inter-trait correlations.—It is conceivable that we have here an explanation of the extremely high average inter-correlation of scores on presumably distinct traits in multiple tests such as Bernreuter's. The correlations between measures of introversion, submissiveness and neurotic tendency may be much greater when the test items are intermingled than when they are grouped into separate tests, because the testee's subjective attitudes are constant throughout the former, but may vary from test to test in the latter*. Analogous results were obtained among tests devized by Smith (1932), Guilford and Guilford (1936), and Stagner (1937); the average inter-correlation of scores on separate measures was 0·24, but this was raised to 0·44 when the items were combined into a single test.

171. Significance of results of factor analysis studies.—We are now in a position to interpret the results obtained from multiple factor

* Stagner (1934, 1937) points out a statistical reason for the exceedingly high Bernreuter inter-correlations. The marks assigned to the various responses for introversion and neurotic tendencies are derived from testees who obtained the highest and lowest scores on separate tests of these two traits. Now though the correlation between the two separate tests may be only + 0·40, yet those testees who get the extreme scores will be much the same in both tests. Hence the two sets of marks, and the Bernreuter test scores, are almost identical with one another. Stagner's explanation is certainly sound, but it will hardly explain the whole of the rise from + 0·40 or less to + 0·93 nor does it account for the other instances cited.

analyses of self-rating tests. It was shown above (§ **149**) that most of the tests, whether they are directed at introversion, or submissiveness, at tendency to fantasy, at parental or sex adjustments, at excitability, depression, worry, shrinking responsibility, or at neurotic tendencies in general, all actually measure one and the same general factor. Although additional factors were needed to account for the whole of the correlations between these diverse tendencies, they were usually of minor importance. A parallel state of affairs was found in Chapter IV, when ratings of others were subjected to factorization ; and it appeared there that the general factor consisted largely of halo. The writer would suggest then that the general factor in self-rating tests similarly consists of the subjective attitudes that we have been considering. Probably it is a highly complex tendency. Those testees who obtain big scores on psychoneurotic, introverted and other traits may be the most conscientious, the most willing to reveal their emotional weaknesses to the experimenter; or they may be the most self-analytic and sophisticated, the most conscious of their weaknesses ; or they may be the most suggestible and neurasthenic. While those who obtain generally low scores may resent the experimenter's curiosity and consciously attempt to draw a favourable picture of themselves or they may be less introspective, more " tough-minded." This interpretation appears to fit all the experimental facts cited above, and to accord with the more speculative generalizations about the subjective state of mind of the testees which we derived from psychoanalytic and other psychological considerations.

172. Validity of self-rating tests.—Thus we are faced with much the same problem as in the previous chapter. It was clear there that the general factor did not consist merely of halo in the minds of the raters, but also represented in part general strength or weakness of character in the ratees. Similarly it is probable that the general factor in self-rating tests does in part correspond to a genuine maladjusted-psychoneurotic-introverted tendency, which is manifested both in overt behaviour and in the judgments of acquaintances. Indeed the two general factors might even turn out to be identical if we could remove halo from the one, and subjective attitude distortions from the other. Again there might be close correspondence between our general factor and Burt's general emotionality (§ **118**), also between the commonly discovered sociality factor (§s **145, 147-149**) and Burt's sthenic-asthenic dimension. But while the separation of halo from ratings did appear to be feasible (§ **113**), the purification of self-ratings is at present quite impracticable. Hence we are not entitled to claim that self-rating tests directly measure the traits by which they are called. Nor can we say whether the second and subsequent factors (obtained after the first has been partialled out) possess a greater validity, or are more free from " self-halo " effects, that the first.

173. Empirical evidence of validity.—We would therefore expect investigations of the validity of self-rating tests to give on the whole

positive, but variable and rather poor, results. The writer has collected from the literature 44 results of comparisons, by 22 investigators, of the tests with ratings of the testees on the same traits by associates; the average correlation is + 0·40, the range — 0·15 to + 0·79. (Here of course both the raters' and testees' halos tend to reduce the agreement). In a detailed investigation of a small group of college students, the writer constructed batteries of tests of various types for measuring a number of personality traits, and so was able to calculate the validity of the different tests. External ratings yielded validity coefficients averaging + 0·60 ± ·09, and five of the self-rating tests described above gave an average figure of + 0·45 ± ·11. Though these figures are low when compared with the validity of intelligence or educational tests, they are as good as, or better than, similar correlations obtained with typical objective tests of temperament and character. And if it is desired to develop batteries of tests which will measure traits as accurately as possible, then there can be no doubt that such batteries should include external and self-ratings. In spite of their defects they do, under good conditions, provide data of partial validity.

174. Other evidence may be derived from comparisons of the test scores of groups with known characteristics. As already mentioned (§ **161**), well and badly adjusted students do show differences if the testing conditions are favourable. Thurstone and Thurstone (1930), Bernreuter (1933a), Stagner (1934) and others have also claimed that extreme scores are genuinely diagnostic of personality disorders; but Landis (1932), Downey (1932), etc. record great discrepancies between the test results and other information about the testees' personalities. Some investigators (cf. Mathews (1923), have shown delinquent children to be somewhat more neurotic than normals. Murray (1932) however discovered no difference, but did obtain higher scores in a group of delinquents considered to be emotionally unstable. Speer (1936) found that none of the Bernreuter scores differentiated significantly between groups of 15–18 year problem cases or delinquents and normals. But he does not describe the kind of co-operation he obtained from his testees. Simpson (1934a) claims a moderate correlation between the Thurstone *Schedule* and recidivism (number of incarcerations) among adult criminals.

175. Results with psychotics and neurotics.—Hunt (1936) summarizes a number of experiments with mental hospital patients, and points out how much depends on the success of the investigator in obtaining candid responses from the patients. We have also suggested that their generally low degree of sophistication and self-awareness tends to reduce their scores (§ **162**). The results are indeed disappointing. In one investigation, Page, Landis and Katz (1934) made up a questionnaire of fifty items which had been judged by competent psychiatrists to be descriptive of the schizoid personality. On applying this to large numbers of dementia praecox and manic-depressive patients and normal persons, the manic-

depressives obtained scores lower by 8 per cent. on the average than the other groups ; but the normals and schizophrenics were almost identical. Nevertheless, a few of the separate test items did differentiate fairly reliably between the three groups. The authors conclude therefore that the *pattern* of responses given by different types of patient may be meaningful, although purely quantitative comparisons based on the *number* of responses are valueless. In another study by Landis, Zubin and Katz (1935), the Bernreuter scores gave no significant differentiation between groups of schizophrenics, manic-depressives, organic psychotics, psychoneurotics and normals. But the Maller *Character Sketches*, a test which was in the first place standardized empirically on abnormal testees, did show somewhat poorer adjustment among the psychoneurotics.

176. Other results indicating validity.—An interesting study by Boynton, Dugger and Turner (1934) demonstrated a greater incidence of symptoms in school children who were taught by teachers with high scores than in children under the charge of more stable teachers. Carter (1935) obtained distinctly greater resemblance between the Bernreuter scores of 55 pairs of identical twins than between those of 74 pairs of non-identicals ; like-sex fraternals were also more similar to each other than unlike-sex pairs. He interprets this as evidence for the partial determination of the test performances by hereditary emotional factors. Studies of family resemblance usually give small positive correlations between husbands and wives, parents and children (cf. Willoughby (1934-36) ; Hoffditz (1934)). Meaningful sex differences occur in tests of dominance and, as reported above (§ **146**) in the Boyd *Questionnaire*.

177. Conclusions and recommendations.—Such results tend to agree with expectations, but are not sufficiently striking to prove that the tests possess great practical value. We are probably justified in concluding that they do measure psychologically significant variables when the testees are adequately motivated to give candid responses. It is likely also that test items which deal with concrete manifestations of personality traits rather than with intimate feelings, and which stress socially acceptable as much as, or more than, unacceptable characteristics, are superior. Almost certainly it is better to approach one trait at a time than to attempt to cover several in a single test ; and factorial analysis techniques might well be employed in deciding on those traits that are sufficiently clear-cut to be tested. Finally, it is very necessary to interpret the test results with caution, remembering that they are always mixed with subjective factors that cannot be fully controlled. Landis (1936) points out that such results are entirely valid as a "self-portrait" ; they represent the picture of himself which the testee wishes to convey to the experimenter. Such a picture may be of considerable value if it is not regarded as a direct measure of some trait, but is compared and contrasted with information about the testee obtained from other sources.

D.—Indirect Measures Obtained from Self-Ratings

(i) *Goodness of Self-Rating*

178. When a testee rates himself, and is rated by acquaintances, on a number of different traits, then the agreement between his own opinions of himself and others' opinions is sometimes assumed to constitute a measure of his self-insight. Should his ratings not merely deviate from those of others, but always be too high on desirable, too low on undesirable traits, we obtain an index of his conceitedness. Allport (1921, 1937) discovered that this self-overevaluation correlates negatively with intelligence and with sense of humour; others also have shown the more intelligent to be more modest. Adams (1927) found that good self-raters are more extraverted, happier, more sociable, etc., whereas good raters of others, though also more intelligent, tend to be more introverted (cf. § **121**). In other words, the person who is interested in others knows himself well, and the person who is interested in himself knows others well. These conclusions are to some extent vitiated by the use of associates' ratings as criteria for the accuracy of self-ratings. An individual who happened to be very unpopular would naturally be rated low on sociability, sense of humour, intelligence, etc., owing to the halo phenomenon; and if he rated himself accurately on such traits his judgments would necessarily deviate from the judgments of his associates, so that he would appear to be a poor self-rater. The writer (1933a) has, however, compared self-ratings with other, more objective measures of certain traits, and his results do confirm Adams's; he also found that good judges of self are not, like good judges of others, above the average in artistic inclinations.

(ii) *The Checking of Extreme Responses, and Variability, in Self-Ratings*

179. Just as in attitude tests (§§ **74–75**), so in personality questionnaires with multiple choice responses, some testees tend to check many more extreme responses than others, both in the neurotic and in the non-neurotic direction. In the Boyd *Questionnaire*, the writer found that the proportion of definite Yes's and No's (as contrasted with Yes?, 0, and No?) ranged from 17 per cent. to 98 per cent. Again, however, this tendency to extremes appears to be unreliable and to possess little psychological significance. Lentz (1934) has studied individual differences in the amount of alteration of responses, when the Bernreuter test is given twice, but, as with attitude tests (cf. § **76**) obtained few results of interest.

VI.—WORD ASSOCIATION METHODS AND INTEREST BLANKS

A. Description of Tests

(i) *Introduction*

180. The chief difference between the following tests and those discussed already is that they do not set out primarily to measure any particular trait, interest or attitude. They might be termed

"buckshot" approaches to personality, in that they fire large numbers of stimuli at the testee, indiscriminately, in the hope that several will reach the mark and will touch off some significant emotional response. Then from these responses a great many traits may be deduced. We will first briefly outline the tests and then show how the responses are treated so as to yield measures of traits and interests.

(ii) *The Word Association Method*

181. First devized by Galton in 1879, the word association method has evolved in many directions, and has been put to a variety of uses. In its commonest form a list of stimulus words is read out, one by one, by the experimenter. To each stimulus the testee is instructed to reply with the first word that comes to mind; not to search about for apt associations but to respond immediately with the first thing he thinks of. Many of the stimuli tend to evoke superficial verbal habits, e.g. "father—mother," "black—white," etc., but a few may touch on the testee's emotional complexes. In those cases he may show hesitation or embarrassment, his response may be delayed, and the response may be some very unusual word. The test therefore provides the experimenter with a lead as to the emotional dispositions around which the testee's life is centred.

182. Further developments of the method.—Many refinements of the technique may be employed. The exact time between stimulus and response may be recorded with chronoscope or stop-watch; reaction times which greatly exceed 2 seconds suggest some mental inhibition. Simultaneous records may be taken of the psychogalvanic reflex; large alterations in electrical resistance are believed to accompany emotional disturbances. Luria and others have obtained fruitful results with an apparatus which records the testee's muscular tonus; this also seems to vary significantly with mental tension. We cannot however deal here with these measures, since our concern is with verbal methods of approach to personality.

183. Stimulus word lists.—Jung's (1916, 1919) list of 100 words is frequently employed, since it contains stimuli likely to evoke a large number of complexes. Kent and Rosanoff's (1910–11) list was selected for a different purpose, and includes words which do not often "call up personal experiences." In some investigations this has been given as a group test, the testees writing their own responses. Here, of course, no timing or other observation of individuals is possible; scoring must be based solely on the content of the responses. Useful word lists for individual application to children at a Child Guidance Clinic have been published by Cattell (1936), and by Burt (1937c). Another, devized by Boyd (unpublished), is used at several Scottish Clinics.

184. Other types of association tests.—In "chain" or "continuous" association tests, a stimulus word is given and the testee is instructed to say every thing that comes to mind in connection with it, or to take each response as stimulus for a fresh association.

Meltzer (1935), for instance, devized a fruitful method of studying children's attitudes to their parents, by getting them to "think aloud" about certain stimuli. He first gave some innocuous practice words like "table, ball, Roosevelt," and then "father" and "mother." The first ten associations to the latter words were recorded.

185. Murray's Thematic Apperception Test.—Another test which seems to possess considerable possibilities in child and student guidance has been called by Morgan and Murray (1935) the *Thematic Apperception Test.* The testee is shown a series of some 10 to 20 pictures in an individual interview, and told to make up a short story to fit each picture, giving free rein to his fantasy. The stories are recorded verbatim. The pictures show a variety of incidents, but in each of them is portrayed some person of the same sex, and about the same age, as the testee. It is found that he tends to project his own needs, sentiments and complexes (conscious or unconscious) into the stories, so that the general themes of the stories correspond in a remarkable manner to his inner emotional life. It is probably easier to obtain good co-operation and revelatory responses by this method than by the more formal word association methods. It falls, however, somewhat outside the scope of this Report, since the stimuli employed are pictorial, not verbal. There are several other analogous tests, e.g. the Rorschach inkblots, where verbal free associations are obtained to a series of meaningless black and white, or coloured, inkblots. With these we will not attempt to deal here.

(iii) *The Pressey X-O Tests*

186. The Pressey *X-O Tests* were originally devized, like word association tests, not as direct measures of any particular traits, but as exploratory instruments (to be applied in group form) by means of which the experimenter could find out what stimuli would arouse various kinds of emotional response in the testees, (cf. Pressey and Chambers (1920), Pressey (1921)). *Form A*, for adults, contains four subtests, each of which consists of a printed list of 125 words, 25 lines of 5 words each. In the first subtest the testees are instructed to cross out (hence the name X–O) all the words that are unpleasant to them; and then to encircle the one word in each line which is most unpleasant. Of the five words in a line, one is presumed to be *unemotional* or a "joker", one refers to *disgust* tendencies, one to *fears*, one to *sex*, and one to *suspicions*; e.g.:—

white drunk choke flirt unfair

In the second subtest each set of five is preceded by one word in capitals; the testee crosses out any of the five which to him are associated with this stimulus word, and then encircles the one word most closely associated. In the third, testees cross out everything which they think is wrong and encircle the most blameworthy in each line. In the fourth they do the same to every word about which they have worried, or felt nervous. As in the first test, the words

are supposed to refer to special types of abnormality, *paranoid, neurotic, schizoid, melancholic* and *hyperchondriacal*, e.g. :

injustice noise self-consciousness discouragement germs

In *Form B*, for children, three similar subtests call for reactions of " wrong, worry and like or interest." Collins (1927) has adapted this *Form* for use in Britain and has published it in full. We will describe later the various methods of scoring.

(iv) *Interest Blanks*

187. Vocational psychologists were early faced with the problem that an individual's estimates of his own occupational interests are often fluctuating and unreliable. He may have very incomplete notions of what the different occupations entail, so that his stated preferences may possess scarcely any value for vocational guidance. " Buckshot " methods were therefore devized, first at the Carnegie Institute of Technology, and later at Stanford University, which record the testee's immediate likes and dislikes for a large number of miscellaneous stimuli, and then deduce his true interests from the total pattern of his responses. We will omit the early interest blanks of Moore, Ream and Freyd, and turn at once to Strong's (1927) *Vocational Interest Blank*, which has superseded them. Their evolution is described by Symonds (1931) and Fryer (1931).

188. Strong's Vocational Interest Blank.—This *Blank* lists :—

I. 100 occupations, e.g. : Actor, Advertiser . . . . Wholesaler, Y.M.C.A. Worker.
II. 54 amusements, e.g. : Golf, Tennis, Chess, Pet Canaries . . . .
III. 39 subjects of study, e.g. : Algebra, Agriculture . . . . Typewriting, Zoology.
IV. 52 miscellaneous activities, e.g. Repairing a clock, Arguments, Saving money . . . .
V. 53 types of people, e.g. Optimists, Pessimists, Foreigners, Cripples, Teetotalers . . . . .

After each of these is printed L I D, and the testee is told to encircle one of these letters to represent like, indifference or dislike, respectively.

VI. Four lists of ten activities or types of people : the testee checks the three in each list he would most like to do or be, and the three he would most dislike. E.g. one list includes :
Caruso, Edison, J. P. Morgan, Pershing, Henry Ford, etc.

VII. Comparison of preferences for 42 miscellaneous pairs of activities, e.g. :
Playing baseball *v.* Watching baseball
Dealing with things *v.* Dealing with people

VIII. A self-rating test of 40 items, e.g. :
Get rattled easily
Am always on time with my work.

Each of these is checked either Yes, ? or No.

It takes approximately 35 minutes to fill in these 404 responses. They may then be treated in a variety of ways so as to throw light on many different interests. Manson (1931) has produced a similar *Blank* for women's vocational interests, Garretson (1930) another for the educational interests of secondary school boys.

189. Studies of children's interests.—We should mention in passing the innumerable investigations, from Stanley Hall onwards, of children's interests, by means of long lists of games, favourite books, etc., which are similarly checked for Like or Dislike. Lehman and Witty (1927) have conducted many studies among thousands of American children with their *Play Quiz*. Furfey's (1928) *Developmental Age Test* includes lists of interests common among boys from 8 to 18 years. Terman (1925) made considerable use of the method in his studies of the differences between highly intelligent or gifted and normal children. He also investigated sex differences, and constructed a provisional classification of play interests into masculine and feminine.

190. Masculine-feminine interests.—Since then Terman and Miles (1936) have constructed a "buckshot" test which is as voluminous as, and even more varied than, Strong's *Blank*, for studying sex differences in adults. This they call the *M–F Attitude Interest Analysis Test*. Its seven subtests include a controlled association test (stimulus words with four responses to choose from, analogous to Pressey X–O); a similar test with inkblot instead of verbal stimuli; a multiple choice test of general knowledge; lists of things and activities which may evoke anger, fear, disgust, pity, blame, etc.; likes and dislikes; and a personality self-rating questionnaire.

B. Qualitative Uses of the Tests

191. The primary use of free word association or continuous association is clinical. The following-up of unusual responses, or of responses which are accompanied by other "complex indicators" (delayed reaction, emotion, etc.) is of more value to the psychiatrist or psychoanalyst than are any of the derived quantitative measures which we describe below. The same might be said of the Pressey *X–O Test*. Collins (1927), Tjaden (1926) and others find the various scores which it yields less useful than the exploration of particular responses in a clinical interview. Shellow (1931) recommends a similar qualitative application of the *Interest Blanks* in a vocational guidance interview. Such usages are hardly scientific, and do not fall within our present purview.

C.—Aggregate Quantitative Scores

192. Word association scores.—In word association, the average reaction time for all the words, or its dispersion, also the average psychogalvanic or kinaesthetic response, and the total numbers of complex indicators have often been investigated as measures of emotionality, or emotional conflict, and the like. Their significance is somewhat doubtful, since they fail to correlate well with any other measures of psychological traits. These also are somewhat outside our scope.

193. X–O affectively score.—The total number of words crossed out in the *X–O Tests* has been termed by Pressey a measure of "Affectivity" or "Richness in emotional association." Its split-

half reliability and repeat reliability over a few days are high, but the latter falls off rapidly over longer intervals (cf. McGeoch and Whitely (1927) ; Thompson and Remmers (1928)). Correlations with other presumed measures of emotionality such as the Woodworth *Inventory*, etc., are generally negligible. Neither does it consistently differentiate delinquent or psychopathological groups from normal persons. On the separate subtests, however, delinquents do on the average appear to have more worries and to regard fewer things as wrong (cf. Courthial (1931) ; Bridges and Bridges (1926)). The author of the test admits that the Affectivity score is a blur of many kinds of response, and does not expect it to have much meaning. It would seem to the present writer that an additional reason for the failure, not only of these measures but of the others described below, may be the total lack of control of the testee's attitudes to the tests. People may interpret worrying, liking, blaming, etc. in so many different ways that no generalized significance can be attached to their scores. There is a great contrast in this respect between the close watch which the clinician keeps on the subjective situation in individual free association, and the haphazardness of the group cross-out tests.

194. Total interest scores.—Strong (1931) sometimes interprets the total number of Likes in his *Blank* as a measure of the breadth or range of interests. The writer, in applying a similar *Blank*, found very large variations in this respect ; some testees gave as few as 35 per cent., others as many as 70 per cent. of L checks. It would seem to him, however, that this measure does not so much represent a trait of " likingness " or " optimism " in the testees, as an extraneous and irrelevant factor, somewhat akin to the checking of moderate or extreme responses in an attitude test (cf. §s **74–75**), or to the varying standards that raters adopt (cf. § **4**). It may be due to the diverse ways in which the testees interpret the meaning of Like, Dislike and Indifferent.

195. Interpretations of Like and Dislike.—Extreme instances of such varying interpretations of *Interest Blank* responses were observed by the writer in applying a short questionnaire on play interests to some 2,000 Glasgow school children. In the younger classes a number of the children had to be assisted individually. Some clearly took Like to mean " socially acceptable," i.e. what their friends usually play with ; others were more concerned as to what would please the teacher. One boy at a Child Guidance Clinic who spent all his time playing with dolls, marked every item in the blank Like, except the item " Playing with Dolls," which was given a Dislike. Presumably that was the best way he could indicate that dolls meant something special to him, or that other children had teased him about them. Though adults are seldom likely to be quite as perverse as this*, yet certainly the meaning of Like and

* Surely however some testees will be tempted to indulge their sense of humour when asked if they like, " Pet canaries, Chopping wood, Pursuing bandits in sherriff's posse, Acting as yell-leader, People with gold teeth, Men who use perfume," and other such items in the Strong *Blank*.

Dislike responses requires much more analysis than it has received so far.

196. Significance of total Likes.—Thorndike (1936) has recently shown concern over this general liking tendency, since it interfered with his studies of types of interests (cf. § **203**). From observations of the behaviour of several of his testees he is inclined to the view that there is a real difference in breadth of interests between those who check many and few Likes, and that this accounts for about one half of the total Likes scores; the other half he ascribes to "constant errors in interpreting and using the L–D scale."

D. Classification of Responses into Types

197. Word association classifications.—Many investigators have attempted to classify the main types of word association responses, and then to find what proportions of a testee's responses fall under each heading. One of the best known classifications is Jung's (1916, 1919). He distinguishes Intrinsic associations (similarity of meaning of stimulus and response); Extrinsic associations (contiguity of stimulus and response in space or time); Clang or sound associations; Miscellaneous, including purely personal, associations, and several sub-classes. Others (e.g. Freyd (1924b)) have classified responses into subjective or egocentric, and objective. Kent and Rosanoff (1910-11) point out pertinently that not only do different psychologists prepare different typologies, but that also they may vary anything up to 35 per cent. in the responses that they assign to any one type. Murphy (1923) has shown that, even with the most careful classification, there is no correspondence between the types to which mental patients are prone and their psychoses. Wells's (1919) attempts to determine the personality correlates of the types by comparison with ratings were similarly unsuccessful. Classifications based more on the content of the responses than on the grammatical nature of the associations seem to be somewhat more significant. Gilliland (1926) used the aggressiveness of responses to words such as "success, danger, enterprize," as one of a battery of tests for aggressiveness. Fisher and Marrow (1934) showed that when elated or depressed moods were induced in their testees by hypnotic suggestion, they tended to produce happier or more melancholy responses, respectively; the latter mood also greatly decreased the speed of response.

198. Classification of continuous associations and fantasies.—Meltzer (1935, 1936) classified children's chain associations to the stimulus words "father" and "mother" under various headings such as types of activities, pleasant or unpleasant tone, degree of dependence or attachment to parents, level of socialization, etc. Different experimenters classifying the same responses agreed in 82 per cent. of instances. Many conclusions as to child-parent relationships were drawn from this material. For example the mentally healthiest attitudes were found on the whole among

children from middle class homes, less healthy from very rich, and worst of all from the poorest homes.

199. Similar treatment is being applied to Murray's *Thematic Apperception Test* (§ **185**). The various stories are classified so as to yield estimates of the strength of a number of traits, complexes, etc. in the testees. The results of this treatment are not yet published, and at present the test is being applied in an almost wholly qualitative manner; the clinician has to interpret from the stories the testee's main personality trends.

200. X–O classified scores.—As mentioned above, the words in two of the Pressey *X–O Form A* subtests are already classified under nine types of abnormality. Allen (1927) and Flugel and Radclyffe (1928) have studied the typological scores (i.e. the number of words under each heading that are crossed out), and find them to be fairly reliable, but to possess poor diagnostic validity when compared with case studies or with answers to questionnaires dealing with the same tendencies.

201. Types of interests.—No systematic classification of interest blank items has been attempted, though the need for it is shown by the frequency with which investigators resort to *ad hoc* classifications in interpreting their experimental results. Thus when Strong (1931) studied the relative Likes and Dislikes of large groups of testees of varying ages, he found marked age differences in items which seemed to fall under such general headings as physical exploits, linguistic subjects of study, administrative occupations, etc. Similarly in contrasting happily married, unhappily married and divorced couples, Johnson and Terman (1935) found the main differences between the groups to lie in logically coherent sets of items, e.g. "uplift interests," "intellectual interests," etc.

202. The present writer prepared an *Interest Blank* with items referring to Spranger's six types of value (cf. §§ **30, 66**). When applied to testees who had also taken an early draft of the *Study of Values*, a contingency coefficient of 0·47 (approximately 0·70 if corrected for attenuation) between the two sets of scores was obtained, showing that the two different techniques do to some extent measure the same general interests. It is doubtful however whether one may use the absolute number of Like and Dislike responses which belong to any one type as an index of the testee's interest in that type, since the general "likingness" factor, which we described above (§§ **194–196**) may affect all such scores. In order to eliminate its influence, the writer employed instead the relative proportions of a testee's responses that belonged to each type.

203. Thorndike (1935, 1936) has done extensive work on the measurement of interests by a blank made up of short sets of items classified under: "Practical activities, Animals, Language, Religion, Children, Gardening," and several other headings. He also found that some testees obtained high absolute scores in all classes, so that the correlations between the various classes were to some extent spurious. However, when the general likingness factor was held

constant by the partial correlation technique, there was still a fair amount of overlapping between the different classes. For instance, "Art, Music, Words and Imagination" classes inter-correlated highly, indicating that they all belong to some wider category. Similarly, "Detail, System, Neatness," and "Animals, Beauty, Children" were linked.

It is clear that much more investigation is needed as to the types of interests which are both self-consistent and distinctive from one another. Strong's (1931), Johnson and Terman's (1935), Thorndike's (1935, 1936), Cattell's (1936) and the writer's classifications are all different, and are all lacking in an empirical basis. We shall see below that multiple factor analysis might be a very useful instrument for studying this problem.

E. Empirical Treatment of Word Association Tests

204. Empirical techniques.—These interpretative qualitative and typological approaches are generally distrusted by American psychologists; hence by far the largest proportion of work on the tests described in this Chapter has been conducted with purely empirical techniques. Much the same method is always used. A long list of verbal stimuli (free association words, X–O words, L I D items, etc.) is applied to a large group of persons, whom we shall call the St-group (Standardization group). In its simplest form, all members of this group possess some common psychological characteristic; e.g. they are mentally normal as distinct from psychotic or neurotic. Their responses to each stimulus are tabulated. The test is now given to the testees (who are quite distinct from the St-group); their responses are compared with the responses of the St-group and are scored according to their degree of resemblance. If they differ widely, the testees are deemed mentally abnormal. Alternatively, the St-group includes two classes of people, e.g. life insurance salesmen and members of other vocations. The main differences between the responses of the two classes are tabulated, and transposed statistically into some convenient form. If now a testee's responses more closely resemble those of the life insurance salesman than those of the other class he is considered to be interested in life insurance salesmanship. This method is, of course, the same as that used by Olson in preparing his rating scale (§§ **124–126**), and by Bernreuter in his *Personality Inventory* (§ **143**).

205. Kent and Rosanoff's work.—It was first applied by Kent and Rosanoff (1910–11) to word associations. They tabulated all the responses of a St-group of 1,000 miscellaneous normal persons to their list of 100 stimulus words, and noted the frequency of each response. When a new testee gives his associations to the same words, the frequency values of all his responses are summed to give a measure of what is called his commonality (if he gets a high score) or idiosyncrasy (a low score). An alternative simpler form of scoring is to note the number of common responses (i.e. the modal or most frequent responses of the St-group), or the number of individual

responses (i.e. responses not listed in the tables owing to the infrequency of their appearance). That the test possesses some value as a measure of mental normality was shown by Kent and Rosanoff's discovery that the 1,000 members of the St-group gave an average of 7 per cent. individual responses each, whereas 247 mental patients gave an average of 27 per cent. each.

206. Kent and Rosanoff's tables are now out of date, since the common responses to their stimulus words have altered considerably since 1910. O'Connor's (1928) tables, based on the responses of 2,000 industrial workers are more often used, (though whether such workers can legitimately be considered a representative normal St-group seems doubtful). The present writer finds a correlation of $+ 0{\cdot}79 \pm {\cdot}014$ between the frequency values of 350 responses selected at random from the Kent-Rosanoff and O'Connor tables. Woodrow and Lowell (1916), using a written instead of an oral form of the test, have prepared tables for children. They found that children's commonest responses were the same as those of adults among only 39 per cent. of the words. These tables are also too old to be valid nowadays. It would, of course, be essential to construct new tables from British St-groups if the test was to be used in this country.

207. Significance of Kent-Rosanoff idiosyncrasy scores.—The meaning of the idiosyncrasy or commonality score has been very variously interpreted. It was originally claimed to show " autistic thinking," but was later widely identified with introversion (Allport (1921) ; Freyd (1924b) ; Guthrie (1927) ; O'Connor (1928) ; Oliver (1930) ; Schwegler (1929) ; Weber and Maijgren (1929)), the supposition being that the extravert would give fewer individual responses. Others have assumed that idiosyncrasy indicates originality (McClatchy (1928)), high intelligence (Olson (1929) ; Wells (1919)), low intelligence (Wheat (1931)), the probable reason for the latter being that duller testees fail to understand some of the words and so give unusual responses. It has also been used as a measure of radicalism as opposed to conservatism (Moore (1925)), and of emotionality (Elonen and Woodrow (1928) ; Laslett and Bennett (1934)).

In actual fact the agreement with self-rating tests or external ratings on introversion are negligible, and correlations with measures of the other suggested traits are so inconsistent as to be valueless. So that our only clue to its meaning is Kent and Rosanoff's result with mental patients, quoted above.

208. Wyman's word association measures of interests.—The next development was carried out by Wyman (1925) and Kelley. S.-groups of children were selected on the basis of teachers' ratings as being keenly or weakly interested in " intellectual, social or activity interests." Their word association responses were tabulated, and differential marks were calculated for each response. Thus, when the test is applied to a new set of children, their responses can be scored for resemblance to those of the intellectual, social or activity interest groups, and measures of these interests obtained. The technique is exceedingly laborious, yet it is objective, and has the considerable

advantage that the testees are not liable to fake their responses, since they can have no conception as to what the test is measuring. The various halo effects found in ratings and self-rating tests are eliminated. Probably, however, halo played a considerable part in the teachers' ratings by means of which the St-groups were chosen, since unduly high correlations of + 0·68 to + 0·80 are found between the three interest scores. Thus Wyman's test measures children's resemblance to groups *reputed to* possess such interests, not their resemblance to unambiguous and objectively defined groups like Kent and Rosanoff's or Strong's. A worse defect is that the St-groups were too small, numbering about 130 for each interest, so that the scores are inconsistent. When the scores of fresh groups of children were compared with similar ratings, the correlations were only + 0·54, + 0·35 and + 0·20.

209. Kelley's word association measures of character.—Still more elaborate and still less successful was Kelley's attempt to measure eight character traits objectively by the same technique (cf. Kelley and Krey (1934)). Teachers' and pupils' ratings were used for selecting St-groups who should be characterized by "Courtesy, Fair play, Honesty, Loyalty to fellows, Mastery, Poise, Regard for property rights, and School drive." When the word association test was applied to a fresh group and their eight scores compared with similar ratings, the correlations averaged only + 0·18. These scores also overlapped very closely with one another (cf. § **117**).

210. It is probable that by taking large enough St-groups and numerous enough word associations somewhat better measures might be obtained. Yet it is difficult to see why anyone should assume that courtesy and the like will be expressed in word associations. These investigators would not attempt to measure arithmetical ability by a test which did not involve arithmetical processes; and yet they expect meaningless methods to yield meaningful results in the field of personality. An examination of some of Wyman's scoring tables reveals scarcely any logical relation between the associations and the types of interest they are intended to measure. For instance, with the stimulus word "Gem", the response "Diamond" scores :—

20 for intellectual, 11 for social and 15 for activity interests.

And the response "Cake" scores :—

3 for intellectual, 9 for social and 12 for activity interests.

Although there is certainly need for objective and scientific methods in the study of personality, it is difficult to believe that this blind empiricism which takes no account whatever of the psychological significance of the test situation and the test responses can yield fruitful results.

F. Empirical Treatment of X—O and Interest Tests

211. X–O idiosyncrasy scores.—In addition to crossing out words the testee is instructed to encircle the one word in each line about which he feels most strongly. Pressey therefore claims to

measure affective idiosyncrasy by finding the number of times a testee's choices differ from the modal words most commonly encircled by a St-group. His original St-groups consisted of 114 College students for *Form A* and 388 students for *Form B*. The first group is much too small, and the second provides an inadequate criterion for scoring the responses of children. Thus it is not surprizing that the split half and repeat reliabilities of the idiosyncrasy scores are low (0·34 to 0·60 according to Flemming (1928); McGeoch and Whitely (1927); Thompson and Remmers (1928)). Collins (1927) has provided useful lists of the commonest responses among 1,500 British children for each sex and age group from 11 + to 14 + years. But the test does not appear to have had any further use in this country. Idiosyncrasy scores are found to be somewhat larger in mentally abnormal and delinquent than in normal groups; thus Collins obtains averages of 43 and 34½ out of 75 among one hundred delinquents and normals respectively (cf. also Bridges and Bridges (1926); Guilford (1926); Tjaden (1926)). But, like the analogous word association measures, they fail to give appreciable and consistent correlations with any other measures of emotional traits (cf. Flemm-ing (1928); Landis, Gullette and Jacobsen (1925)).

212. Chambers (1925ab) and Weber (1932) have worked out differential score values, analogous to Wyman's, from St-groups of scholastically superior and inferior students, and from boys of different ages, so that the test may be scored for scholastic interests and for maturity of emotional responses. The former measure gave no agreement with scholastic success when tried out on a fresh group (cf. Thompson and Remmers (1928)). Weber's *Emotional Age Scale*, however, which incorporates much fresh material such as lists of liked or disliked books and games along with the *X–O Form B Tests*, does show better evidence of consistency and validity. Though it has not been widely applied, it might provide a useful supplement to tests of intellectual development. Furfey's *Developmental Age Test* is similar (cf. § **189**).

213. Vocational interest scores.—Vocational interest blanks are always standardized by reference to the responses of St-groups of persons engaged in particular vocations. The early tests of Freyd (1924b) and others were too short, and were applied to insufficiently large groups; hence the responses found to be typical of groups of salesmen and mechanics failed to differentiate effectively between other salesmen and mechanics. The same criticism applies to some of Strong's vocational groups, though most of them are very large. They should include about five hundred members of a vocation if the test is to attain an adequate consistency. Strong's *Blank* (1927) may now be scored for the resemblance of a testee's responses to the responses of some thirty different vocational groups, including Architects, Artists, Boy Scout Masters, Certified Public Accountants, Chemists, Doctors, Farmers, Journalists, Policemen, Psychologists, Teachers and Vacuum Cleaner Salesmen. Needless to say his test cannot legitimately be employed in this country, since it is

far from likely that the likes and dislikes of Californian and British vocational groups will be sufficiently similar. Manson's (1931) *Blank* may be scored for ten common women's vocations ; Garretson's (1930) for three main types of secondary education, Academic, Commercial and Technical.

214. Derivation of vocational interest scores.—We will outline briefly the method by which Strong determines the various interest marks for each item. Several different statistical techniques have been employed by different compilers of empirical tests (e.g. Allport (1928) ; Kelley and Krey (1934) ; Flanagan (1935) ; Humm and Wadsworth (1935)). But Strong finds the following simple technique as effective as any. Suppose we wish to calculate the marks for interest in Personnel Management which are to be assigned to the item : ACTOR L I D. The percentages of the St-group of personnel managers who check each response are noted, also the percentages of members of all other vocational groups, with the following result :—

	L	I	D
Personnel Managers	49	38	13
All others	38	35	27
Difference	+11	+3	−14

The differences are then transposed into somewhat smaller figures, and the final marks for this item are + 2, + 1 and − 3, respectively. Other marks are similarly determined for each of the 1,200 odd possible responses, for each vocational group. An individual testee's score for a vocational interest is the sum of his + and − marks, as derived from these groups. Scoring a single testee's blank on 30 vocational interests takes several hours, unless a Hollerith machine is available. If the testee's score falls within the range of scores obtained by 75 per cent. of the St-group, he is given an A-rating for that vocational interest ; if it is within the range of the lowest 25 per cent., he is given a B-rating ; if it is lower than the scores obtained by any member of the St-group, a C-rating. A candidate for vocational guidance who takes the test is advized only to enter those vocations on which he receives an A-rating.

215. Reliability and validity of vocational interest scores.—Strong (1935) has shown that the reliability of vocational interests in adults over a five-year period is represented by a correlation of 0·75 (0·84 when corrected for attenuation). Naturally, they are somewhat less stable in adolescents. Persons who are tested before they take up their careers (but not told their results) do tend spontaneously to enter vocations which correspond well with their test scores. Those who later forsake a vocation are found to have had lower scores than those who stay in it. And in at least one vocation, life insurance selling, there is some correlation between success at the job and possession of an A, B or C interest rating. Clearly then the test does work practically.

216. In other researches Strong (1931) has obtained results from St-groups of different ages from 15 to 55 years, so that the test may be scored for "interest maturity," and from the two sex groups. Probably however a more valid test of sex-resemblances is provided by Terman and Miles's (1936) *M–F Test*. We will summarize briefly some of their work, and leave on one side the many applications of empirical techniques to other fields.

217. Terman and Miles's investigation of sex differences.—Very large numbers of items which seemed likely to differentiate the sexes were tried out on men and women, and those which were found to be significantly related to sex were incorporated in the two parallel forms of the *M–F Test*; + (male) and −(female) marks were determined for each response to each item, but in the final version every item is weighted equally so that the scoring process is fairly easy. The range of total scores found among men is + 200 to − 100, among women + 100 to − 200, the medians being + 52 and − 70. Thus there is a good deal of overlapping, indicating that some men possess interests, attitudes, etc. more feminine than the average woman, and vice versa. Highly meaningful differences are found between the average scores of different groups; e.g. college athletes are more masculine than the norm; persons outstanding in their career, especially in engineering or scientific careers are also superior; artists and theological students are below the norm, and passive male homosexuals tend to approach the feminine median. Among women similar differences occur, the most domestic types yielding the lowest scores. These and other quantitative results, combined with qualitative interpretation of the best differentiating items, provide a mass of interesting data on the psychology of the sexes.

218. Discussion of interest blanks.—We may conclude, then, that empirical methods do give decidedly better results when applied to Likes and Dislikes, or to material which was logically selected as was Terman and Miles's, than they do with word association and cross-out tests. We also find in Strong's and Terman's tests (unlike the Wyman-Kelley tests) that there is a fairly clear meaningful relation between the marks for many of the responses and the interests to which they have been proved, empirically, to correspond. The highest marks undoubtedly occur in responses which one might expect to be characteristic of the various vocational or sex groups.

219. Actually there is a slight disadvantage in this, since it permits a testee to fake his responses if he is aware of the object of the tests. Steinmetz (1932) has demonstrated that not only can a testee obtain an A or B rating on any vocational interest which he is intentionally simulating, but that also his scores on half the other interests are seriously distorted by this faking. And Kelley, Miles and Terman (1936) discovered that a man could, if motivated to do so, make himself out considerably more feminine than the average woman, or vice versa. Nevertheless it must be allowed that these tests are much less liable to faking than are most personality

questionnaires, that testees are seldom likely to guess their object, and that it is relatively easy to get good co-operation and frank responses. The tests are in fact more objective than any we have described, with the exception of the direct observation and time-sampling techniques (§ **95**).

220. We are however entitled to ask, what are these tests measuring? The various resemblance scores do not represent interests in the usual psychological sense of the term; though it may well be that they are of more practical value than any of the purer measures of interests that have been devized. Nor does the *M–F Test* correspond at all closely with ratings on masculinity-femininity, one good reason for this being that the raters are quite unable to agree on the definition or application of this characteristic. After studying the results of many investigations in this field, the writer is strongly impressed by the frequency with which the investigators relapse from their Behavioristic empiricism. It is not usually the objective scores but the responses to particular items, or to subjectively classified types of items, which seem to reveal most about the psychology of personality. Possibly then, although these methods are so ingenious and extensive, they are still rather barren from the research standpoint; and there still seems to be room for psychologists who try to understand, and not merely to measure, human traits. The dangers of relying solely on empirical treatment are well illustrated by an experiment of Burnham and Crawford (1935). They obtained a set of purely chance scores on the Bernreuter *Inventory* and Strong *Interest Blank*, by throwing dice to decide each response. When the results were scored in the usual way it was found that the dice possessed most of the characteristics of an emotionally maladjusted boy scout master or journalist.

Finally, we would point out that these empirical methods are tremendously laborious, and so consuming of time and money that, although parallel tests would be decidedly useful in this country, it is possible that no country other than the United States will be able to afford them. We hope then that some day there may be devized alternative methods which, while not less scientific, may take more account of psychological considerations, and perhaps at the same time be less roundabout and less wasteful.

G. Multiple Factor Analysis of Vocational Interests

221. One of the lines of advance should certainly be through factor analysis. It is obvious that there must be great overlapping between many of the scores for different occupations. Theoretically we should be able to reduce these to not more than half a dozen main types of vocational interest; then instead of scoring a testee separately for each vocation, we could deduce from the patterning of his six factor scores his interests in an indefinite number of more specific fields. Moreover, better tests of such factors might be devized, which could be more economically scored.

222. Gundlach and Gerum's investigation.—From inspection of the interest inter-correlations, Gundlach and Gerum (1931) deduced five main types; Social, Intellectual, Technical, Creative, and Physical Skill interests, together with several sub-types. These were found to be only moderately self-consistent and discrete from one another. It was shown that each specific vocation possessed a distinctive pattern of interests on these types. A more exact technique of analysis is provided by factorization.

223. Thurstone's investigation.—Thurstone (1931b) himself factorized the scores of 287 testees on 18 interests, and found four main factors which seemed to correspond in a general way to interest in science, in language, in business and in people*. A few of the scores however still showed prominent specific factors, i.e. their variance was not fully accounted for by these four factors. Hence the next step, the prediction of many different scores from combinations of a few factor scores, has not yet been attempted. It is interesting to note the close correspondence between these empirically determined factors and the four types—theoretical, aesthetic, economic and social—which Spranger (1928) derived from logical and intuitive considerations. It reinforces our suggestion that more rapid progress might be made through the co-operation of "arm-chair" psychology with experimental testing and statistics.

224. Other factorial studies.—Strong (1934) repeated the analysis, including six more interest scores, and obtained a distinctly different set of five factors, which were less easy to identify. A partial explanation of this result may be his failure to rotate his axes appropriately, but it may also indicate (what we pointed out in § **38**) that the factors are not universal in that they are limited by the particular set of measures which are factorized. A further difficulty is suggested by Carter, Pyles and Bretnall's (1935) analysis of 23 interest scores, obtained from 133 youths of medium age 16½ years. Here the factors resemble but do not coincide with either Thurstone's or Strong's, and they probably reflect the difference between the organization of interests among adolescents and adults.

VII. DISCUSSION AND CONCLUSIONS

A. Verbal Tests and Ratings

225. We have described a considerable number of, often very ingenious, techniques which have been devized for researches on personality and social psychology; and those specifically mentioned or illustrated are, it should be remembered, but a small, though we hope representative, sample of the techniques available. In spite of their diversity, both of form and of purpose, several general conclusions may be drawn which appear to apply to all of them.

* One incidental result worth mention was that the Psychologist's interest score gave a zero correlation with the "Interest in People" factor. Thurstone regards this as quite probably true.

226. Contrast between verbal personality tests and tests of abilities.—First it is clear that none of these tests can claim to measure psychological variables such as traits, attitudes or interests with the same degree of objectivity and accuracy that are achieved by tests of abilities. At least one reason for this is obvious, namely that the criterion of ability in any line is some performance which makes its mark on the physical world, and which may be readily recognized, identified and measured. The individual with high mechanical ability is one who can do something which we would universally agree to be mechanical in nature*. Whereas behaviour of an emotional or conative character cannot be so easily specified; there is considerable room for disagreement as to whether or not such and such an action (either bodily or verbal) is or is not an expression of introversion, of aesthetic taste, of trustworthiness, etc. We have seen how difficult it is adequately to define the content of an attitude like radicalism (§ **39**); still more ambiguity is likely to occur among character and temperamental traits, which often have little or no effect on the individual's physical environment, but are expressed only in his personal reactions and in the impressions he makes upon others.

227. Difficulty of quantification of traits.—Not only do affective and conative traits possess less objective "subject-matter" than abilities, they are also less amenable to consideration as uni-dimensional variables. Certainly different people differ widely in their sociability, tolerance, etc.; but it seems more artificial to grade them according to the amounts of a trait that they possess than it does to quantify their amounts of an ability. A quantitative change appears to involve also a marked qualitative change. For example, an extremely methodical person is not like a fairly methodical person only more so; the former's trait seems to take on a new shade, and to be so bound up with the rest of his personality that we find it extremely difficult to compare him with the latter person simply in terms of more or less. A further conclusion which was reached in discussing radicalism—that it is a complex pattern rather than a single continuum—holds good among the other traits and interests with which we are concerned. It would hardly be possible to build up a science of human nature without such quantification of differences between individuals or between groups, and without measures of traits *as if* they were linear variables; yet we must admit that all our measurements are at present, and are likely to remain, relatively abstract, inaccurate and awkward as descriptions of the rich complexities of our psychological material (cf. Allport (1937)).

When therefore the results of investigations reported in the previous five chapters appear disappointing, or progress seems slow in comparison with the efforts expended, it may simply be that our

* Many abilities, e.g. managerial, teaching, etc., do not, of course, produce obvious physical effects; and these are quite as hard to define and measure as are personality traits.

instruments are at present inadequate or inappropriate for the problems to which they are applied. Instances are afforded by Landis's observation that differences do exist between mental patients and normals in questionnaire responses, though these differences are obscured in the total questionnaire scores § (**175**) ; and by the writer's demonstration that more can be deduced from the features and external appearance than the results of ordinary rating experiments would allow (§ **100**). At the same time there is so much evidence of improvements in these instruments (e.g. the substitution of graphic for numerical rating scales, the studies of the form of questions in attitude scales and personality questionnaires, etc.) that we certainly cannot set a limit to the future possibilities of quantitative methods in this field. Moreover, the poor tools at our present disposal have already revealed much which was unrecognised in the days of purely qualitative observation and interpretation of human phenomena ; thus there is good reason to hope for greater advances to come.

228. Advantages and disadvantages of verbal over behavioural tests.—It is largely owing to the indefiniteness of the behavioural content of traits, attitudes and interests, that verbal methods have been so extensively developed. Words are actions in miniature. Hence by the use of questions and answers we can obtain information about a vast number of actions in a short space of time, the actual observation and measurement of which would be impracticable. Not infrequently, also, a trait is more directly expressed in subjective feelings, or in the effects it makes on other people, than it is in concrete actions, so that the verbal opinions of the possessor of the trait or of his acquaintances are the only possible sources of information about that trait. But unfortunately, words, though originally the correlates of actions, are much more generalized and abstract than actions ; they are interpretations rather than descriptions, and are fraught with ambiguities. The questions which are put in attitude scales, in ratings, personality inventories and interest blanks, often mean different things to different individuals, and the responses are similarly equivocal. Our pious hope that the inclusion of a large number of questions in each test, or that the use of analytic ratings filled up by several raters will cancel out the errors inherent in any one question and answer or rating, is seldom justified. For there is ample evidence in the halo phenomenon, the corresponding phenomena in self-rating tests, and the general "likingness" factor in interest blanks, that such errors may be constant rather than variable.

229. Need for more careful psychological study of the tests.—In our opinion a major line of advance will consist in a more careful analysis of these complex subjective factors bearing on the interpretation of test questions and the psychological significance of test responses. A useful start might be made with a thorough introspective study, in the German tradition, of the mental processes involved in rating, in answering Yes *or* No, Like *or* Dislike, etc. The

trend of thought in American investigations has always been opposed to such subjective approaches. Symonds (1931) is representative of the majority when he states, in effect, that subjective reactions to particular test items need not concern us because the significance or validity of the test as a whole should always be determined empirically ; that the resistances and inhibitions aroused by many of the tests do not matter to the psychometrist, since the psychological variables which such tests measure should be established by objective correlational comparisons with other variables rather than by subjective speculations or " arm-chair " considerations. Yet it may be precisely this neglect of subjective analysis which is responsible for the poor empirical validity of so much of the work, and for the apparently slow speed of advance. Certainly "arm-chair " psychology alone will not solve our problems, but it is, and always has been, an essential precursor to the most fruitful experimental research.

230. Difficulties due to variability and ease of simulation of personality traits.—A further fundamental difference between an ability and an emotional characteristic is the greater variability in the latter from time to time. The level which we maintain in performing a series of arithmetical problems is much more constant than our moods. Even the habitually depressed person has at times been elated, the happy-go-lucky man is sometimes cautious. Thus it is possible for us to simulate almost any trait or attitude with a fair degree of success. In everyday life we normally adapt our personalities to some extent to the company we are in, for instance, showing different characteristics at work, at home, and at a party. It is only natural then that the testee should adapt his personality into a mould which he regards as appropriate to an experimental test situation. Often in everyday life a clever observer can penetrate through our disguises and see whether or not our moods and sentiments are genuine. But the personality test can hardly lay claim to similar insight ; it can usually only record the testee's words or actions at their face value. And when a test does not even record actions, but only descriptions or interpretations of actions, the testee may falsify his responses with the greatest of ease. All the tests and ratings that we have described are open to this admittedly serious defect, though those described in Chapter VI (word associations and interest blanks) are relatively free from it.

231. Complete dependence of the tests on the testee's *Einstellung*.—It is clear then that the methods with which the Report deals are likely to be rendered valueless should the testees or raters have any motive for falsification ; and that it is seldom possible to detect when such falsification has occurred. To some extent the object of the tests, i.e. the traits or attitudes at which they are aimed, can be hidden, but not sufficiently for them to be employed for vocational selection, or for other purposes where personal advantage enters. Such disguise is still less possible with ratings, unless Olson's empirical technique be used. We are therefore entirely dependent upon the

good faith and the co-operation of the testees or raters. Either they must be convinced that candid responses will be advantageous to them personally, or else persuaded that the investigation is of scientific value and that candidness will not be in any way disadvantageous.

There is no reason to suppose that deliberate falsification played a large part in any of the investigations we have described. But unwitting distortions, whose effects are similar, are likely to be ubiquitous. The average rater always tends to mark his friends too high on desirable traits because he quite genuinely regards them as superior; the neurasthenic attributes to himself numbers of emotional weaknesses which the "tough-minded" rejects in all sincerity. The term personality derives etymologically from "a mask." It embodies the notion that we are playing a part in some drama. And we seldom realize the extent to which the traits and attitudes that we display to the public gaze, or express in these tests, or admit to ourselves, are assumed. Whether even the psychoanalyst can penetrate to the true inner core of our temperament or character is doubtful; certainly no test can do so.

232. Value of the results.—The conclusion follows, not that tests and ratings are useless, but that their results must always be interpreted in the light of their origins; that the probable subjective attitude of testee or rater should be taken into account in deciding on their significance. Their "fictional" nature may be a merit rather than a defect. For instance, the psychologist or psychiatrist who possesses other sources of information about an individual who has answered a personality inventory may, by comparing the latter with the former, be able to discover the individual's inhibitions, self-deceptions and ego-ideals, which are of vital importance in diagnosing and treating him. Discrepancies between tested attitudes and observations of behaviour, either of individuals or of groups, may be equally revealing. From this point of view, qualitative analysis of the answers to particular test items may be more valuable than the quantitative aggregate scores. Nevertheless, the scores alone almost always do show some objective validity when correlated with other criteria (cf. **§s 59, 110, 173, 215,** etc.), especially when careful attention is paid to the conditions under which they are obtained, and when the traits or attitudes tested are not of a highly intimate or personal character. So that though it is unwise to take them at their face value as direct measurements of these traits, they may justifiably be regarded as partial indicators. For instance, they may be combined with other tests (liable to other types of error), or with other ratings, in a composite battery which will measure the traits with quite a high degree of validity (cf. Vernon (1934a)).

233. Future research.—Future progress will depend then mainly on the better adaptation of the instruments to the fields in which they are already being applied. As suggested above, introspective and qualitative analyses of the test situations are needed; but there

is already room for a large amount of controlled experimental research on the form of the tests, wording of items, influence of instructions, etc. The mere piling up of more tests, and more tabulations of group differences in test scores or test items—of which so big a proportion of American investigation consists—is of very little value. More intensive studies of particular social-psychological problems by these verbal tests and by other techniques (observational and clinical) would be especially fruitful in throwing light on the advantages and disadvantages of the verbal tests.

234. We are inclined also to recommend very thorough examination of a verbal method which received only incidental mention in the Report, but which might prove to be relatively free from most of the difficulties we have just described: that is, the giving of ratings, either on general traits or attitudes or on more specific behaviour characteristics, by a psychologist or psychiatrist on the basis of an interview with a ratee or with acquaintances of the ratee. The ratings assigned to vocational guidance candidates at the National Institute of Industrial Psychology (§ **103**) approximate to this type; the *Vineland Social Maturity Scale* (§ **88**) is an even better instance, in view of the objective nature of its component items. In the interview the ratee, or his acquaintances, would be asked to supply concrete information as to his thoughts, feelings, and actions which are pertinent to the trait or item being rated; and the psychologist would judge whether this information indicated a high, medium, or low rating. Misunderstandings of the test material would be enormously reduced by this method, and witting or unwitting falsifications or disguises might be penetrated by an experienced examiner. The examiner could not, of course, hope to be entirely immune himself from halo and other prejudices (e.g. a Freudian and an Adlerian might give different ratings to the same person), But the extent of such errors could readily be investigated experimentally; and Doll's (1936 a, b) results suggest that it may be very small. When different examiners, often interviewing different informants, filled in the *Social Maturity Scale*, the correlations between the scores which they assigned to the same patients amounted to about + 0·90.

B. Multiple Factor Analysis

235. Uses of factor analysis.—Although we have advocated above a more careful study of the methods by which verbal data are obtained, this does not preclude further analysis of the data itself, by factorization or other statistical techniques. We should indeed remember that no amount of statistics can improve on data which are inaccurate in the first place, and that the factors extracted from ratings or test scores are as impregnated with errors as are these ratings or scores. Nevertheless, the factorization of personality inventories was of considerable assistance in determining the nature of the errors (§ **171**), and statistical treatment of ratings showed promise of enabling us to separate off the halo effect in ratings (§ **113**). Elsewhere it was found that factorial analysis was useful

in classifying overlapping variables such as the nineteen tendencies of the Boyd *Questionnaire* (§ **146**), and the twenty or thirty vocational interest measures derived from the Strong *Blank* (§§ **223-224**); also in analyzing poorly defined psychological conceptions such as introversion-extraversion into distinctive components (§ **147**). In other words, factorial techniques constitute a powerful instrument for the generalization and systematization of test results. It is impossible to study personality or social phenomena without classifying and analyzing, and we are far too apt to do so without any scientific backing. For example, we distinguish types of interests or of abnormal mental states, and assume that these are discrete from one another, or that they are accompanied by various personality traits, mainly on the grounds of subjective generalizations. And we often imply the existence of general traits or factors, for which there is no real evidence, when we make predictions about people, either at Clinics, in vocational guidance, or in everyday life. By means of factor analysis we should ultimately be able to do all this objectively.

236. Limitations of factor analysis.—In introducing the topic of factor analysis (§ **61**) we implied, as do most statistical psychologists, that such analysis would reveal the underlying structure of the personalities to whom our tests are applied. But we can see now that this claim is somewhat presumptuous, and that it would be safer to regard analysis merely as revealing the logical structure of the applied tests. Kelley (1935) and others do indeed talk of isolating the unitary traits or basic elements of personality by means of these techniques and hope eventually to establish a relatively small number of independent dimensions or elements, in terms of which any test may be classified, and any personality completely described. Now this conception of personality as a compound of a few elements is very attractive to the psychometrist, who wishes to measure people as economically as possible; but it is a conception for which not the slightest justification may be derived from biology, general psychology or psychoanalysis. As Burks (1936) shows, personality is more likely to be made up of a multitude of complexly inter-related dispositions than of a few discrete traits and abilities*. Nor do the results so far obtained by factorial investigations lend much support to the geometrical view.

We have already seen that the factors extracted are governed by the particular set of tests that the investigator chooses to apply (§§ **38, 224**). It is no longer true, as it was when Thurstone's technique was first put forward, that the insertion of a few additional tests into the set, or the omission of a few, entirely alters the resultant factors; for by rotation of axes we can maintain the composition of the factors approximately constant. Nevertheless, it is obvious that the factors can only cover those facets of personality

* The factorist's offer to supply inter-correlated factors instead of the usual independent (orthogonal) axes would not, of course, help at all to close the gap between these two contrasted views of personality.

which are represented in the test battery ; hence their universality is limited by the comprehensiveness of the sampling of human traits. Similarly, in the field of interests, the main dimensions will inevitably vary with the particular measures included, until such time as a method is devized for specifying and measuring the whole range of interests. Kelley (1935) has realized this point and has therefore attempted to prepare a *complete* classification of vocational interests.

We have seen also, that the available tests cannot be accepted as direct measures of personality, owing to their distortion by subjective attitudes, etc. Hence, although a fair degree of unanimity is found among different factorizations of ratings (§ **119**), or different factorizations of self-ratings (§ **149**) and attitude tests (§ **65**), there is no direct agreement between the results from the two types of data. The third main type, namely objective measures of conduct, has been too little explored as yet for us to determine whether factors similar to "*w*", general adjustment, sociability, radicalism, etc., are manifested in behaviour. Clearly then none of the factorizations we have described can claim to have disclosed the real elements of personality.

237. Conclusion.—More fundamental is the objection that, while the test inter-correlations are consistent with the extracted factors, they do not prove that these are the only possible factors. It is now generally admitted that an infinite number of different factorizations of the same set of variables is possible (cf. Thomson (1935)) ; and that as their relative merits cannot be decided solely by mathematical considerations, the most suitable alternative must be selected on logical grounds. Holzinger (1936) states that : " What the factorist seeks is the simplest, most parsimonious, and most useful pattern for the interpretation of the underlying variables." Factors are " a possible way of thinking about mental traits." We would conclude then that they should not be regarded as faculties or entities existent in the personalities tested, but as convenient descriptive categories, which enable us to generalize and simplify our test results, and to make predictions about people with a maximum degree of efficiency.

Although then, there is no obligation to accept any investigator's factors as representing the true dimensions of human nature (even *g* is, from this point of view, a convenient fiction), yet it would make for more rapid progress if different investigators would take more account of the work of their predecessors, by choosing for their main axes factors that are already widely established. If, in future studies, more care is taken in obtaining comprehensive sets of tests, chosen on the basis of some logical analysis of the field, and in connecting up as many clusters of these tests as possible with previously determined clusters, we may hope before long to achieve a fairly complete classification of all our psychological measuring instruments which would be of the utmost value in many branches of pure and of applied psychology.

ACKNOWLEDGMENTS

In the preparation of this Report, I have received valuable assistance from numerous sources. Much of the material was originally gathered during my tenure of a Rockefeller Fellowship in Social Sciences at Yale and at Harvard Universities in 1929–31. Among the many persons who provided facilities at that time for my investigations of personality tests, I would especially mention Professor M. A. May, the staff of the Yale Personnel Bureau, and Professors G. W. Allport and H. A. Murray. Further stimulating advice was obtained at Harvard and Columbia Universities during a shorter visit in 1937, from Professors Allport, Cantril, and others. Helpful criticisms of earlier drafts of the Report were given by Professor F. C. Bartlett and Professor C. L. Burt. None of these persons, however, should be held responsible for the views expressed above. I am grateful, also, for the co-operation and patience of students in America and at Jordanhill Training Centre, Glasgow, who acted as subjects in the experiments referred to in several sections of the Report.

Acknowledgments for permission to reproduce illustrative items from published tests is made to the following :—

University of London Press—R. B. Cattell's *Projection Test.*

Dr. E. A. Doll, The Training School, Vineland, New Jersey—*The Vineland Social Maturity Scale.*

Professor D. A. Laird, Colgate University, Hamilton, N.Y.—*Personal Inventories B2* and *C3.*

Professor T. F. Lentz, Washington University, St. Louis Mo.—*C. R. Opinionaire.*

Houghton Mifflin Co., Boston, Mass.—Allport's *A–S Reaction Study* and Allport-Vernon's *Study of Values.*

Stanford University Press, Calif.—Strong's *Vocational Interest Blank,* Bernreuter's *Personality Inventory,* and Willoughby's *Emotional Maturity Scale.*

Bureau of Publications, Teachers' College, Columbia, N.Y.—Watson's *Test of Public Opinion,* and Maller's *Character Sketches.*

University of Chicago Press, Ill.—Thurstone and Chave's *Attitude toward the Church Scale.*

C. H. Stoelting Co., Chicago, Ill.—Woodworth's *Personal Data Sheet,* and Pressey's *X–O Test.*

REFERENCES

ADAMS, H. F. (1927) : The good judge of personality. *J. abnorm. soc. Psychol.*, **22**, 172.

ALEXANDER, F. (1934) : Evaluation of statistical and analytical methods in psychiatry and psychology. *Amer. J. Orthopsychiat.*, **4**, 433.

ALLEN, E. A. (1927) : Temperamental tests *Brit. J. med. Psychol.*, **7**, 301.

ALLPORT, F. H. (1934) : The J-curve hypothesis of conforming behaviour. *J. soc. Psychol.*, **5**, 141.

ALLPORT, F. H., and ALLPORT, G. W. (1921) : Personality traits : their classification and measurement. *J. abnorm. soc. Psychol.*, **16**, 6.

ALLPORT, F. H., and HARTMAN, D. A. (1925): The measurement and motivation of atypical opinion in a certain group. *Amer. pol. Sci. Rev.*, **19**, 735.
ALLPORT, G. W. (1928): A test for ascendance-submission. *J. abnorm. soc. Psychol.*, **23**, 118.
—— (1929): The composition of political attitudes. *Amer. J. Sociol.*, **35**, 220.
—— (1935): Attitudes. *A Handbook of Social Psychology* (ed. C. Murchison). Worcester, Mass.: Clark University Press. P. 798.
—— (1937): *Personality: A Psychological Interpretation.* New York: Holt.
ALLPORT, G. W., and CANTRIL, H. (1934): Judging personality from voice. *J. soc. Psychol.*, **5**, 37.
ALLPORT, G. W., and ODBERT, H. S. (1936): Trait-names: a psycholexical study. *Psychol. Monogr.*, **47**, No. 211.
ALLPORT, G. W., and VERNON, P. E. (1931): *A Study of Values.* Boston, Mass: Houghton Mifflin.
ANDERSON, V. V. (1929): *Psychiatry in Industry.* New York: Harper.
ARGELANDER, A. (1937): The personal factor in judging human character. *Character & Pers.*, **5**, 285.
BARRY, H. (1931): A test for negativism and compliance. *J. abnorm. soc. Psychol.*, **25**, 373.
BERNREUTER, R. G. (1931): *Personality Inventory.* Stanford University Press.
—— (1933a) The theory and construction of the personality inventory. *J. soc. Psychol.*, **4**, 387.
—— (1933b): Validity of the personality inventory. *Person. J.*, **11**, 383.
—— (1933c): The measurement of self-sufficiency. *J. abnorm. soc. Psychol.*, **28**, 291.
BLOCK, V. L. (1937): Conflicts of adolescents with their mothers. *Ibid.*, **32**, 193.
BOGARDUS, E. S. (1933): A social distance scale. *Sociol. & soc. Res.*, **17**, 265.
BOYNTON, P. L., DUGGER, H., and TURNER, M. (1934): The emotional stability of teachers and pupils. *J. juv. Res.*, **18**, 223.
BRIDGES, J. W., and BRIDGES, K. M. B. (1926): A psychological study of juvenile delinquency by group methods, *Genet. Psychol. Monogr.*, 1926, **1**, No. 5.
BRIDGES, K. M. B. (1931): *Social and Emotional Development of the Pre-School Child.* London: Routledge.
BURKS, B. S. (1936): Personality theories in relation to measurement. *J. soc. Psychol.*, **7**, 140.
BURNHAM, P. S., and CRAWFORD, A. B. (1935): The vocational interests and personality test scores of a pair of dice. *J. educ. Psychol.*, **26**, 508.
BURT, C. L. (1915): General and specific factors underlying the primary emotions. *Rep. Brit. Ass.*, 694.
—— (1919): Facial expression as an index of mentality. *Child Study*, **12**, 1.
—— (1937a): Methods of factor-analysis with and without successive approximation. *Brit. J. educ. Psychol.*, **7**, 172.
—— (1937b): Correlations between persons. *Brit. J. Psychol.*, **28**, 59.
—— (1937c): *The Subnormal Mind* (2nd ed.). Oxford University Press.
—— (1938): The analysis of temperament. *Brit. J. med. Psychol.*, **17**, 158.
BURT, C. L., GAW, F., et al. (1926): A study in vocational guidance. *Rep. industr. Fatig. Res. Bd., Lond.* No. **33**.
CADY, V. M. (1923): The estimation of juvenile incorrigibility. *J. Delinqu. Monogr.*, **2**.
CANTRIL, H. (1932): General and specific attitudes. *Psychol. Monogr.*, **42**, No. 192.
CANTRIL, H., and ALLPORT, G. W. (1933): Recent applications of the Study of Values. *J. abnorm. soc. Psychol.*, **28**, 259.
CARLSON, H. B. (1934): Attitudes of undergraduate students. *J. soc. Psychol.*, **5**, 202.
CARTER, H. D. (1935): Twin-similarities in emotional traits. *Character & Pers.*, **4**, 61.

CARTER, H. D., CONRAD, H. S., and JONES, M. C. (1935) : A multiple factor study of children's annoyances. *J. genet. Psychol.*, **47**, 282.
CARTER, H. D., PYLES, M. K., and BRETNALL, E. P. (1935) : A comparative study of factors in vocational interest scores of high school boys. *J. educ. Psychol.*, **26**, 81.
CASON, H. (1930) : An annoyance test and some research problems. *J. abnorm. soc. Pyschol.*, **25**, 224.
CATTELL, R. B. (1933) : Temperament tests. *Brit. J. Psychol.*, **23**, 308.
—— (1936) : *A Guide to Mental Testing*. University of London Press.
CHAMBERS, O. R. (1925a) : Character trait tests and the prognosis of college achievement. *J. abnorm. soc. Psychol.*, **20**, 303.
—— (1925b) : A method of measuring the emotional maturity of children. *Pedagog. Sem.*, **32**, 637.
CHANT, S. N. F., and FREEDMAN, S. S. (1934) : A quantitative comparison of the nationality preferences of two groups. *J. soc. Psychol.*, **5**, 116.
CHANT, S. N. F., and MYERS, C. R. (1936) : An approach to the measurement of mental health. *Amer J. Orthopsychiat.*, **6**, 134.
CHI, P. L. (1937) : Statistical analysis of personality ratings. *J. exp. Educ.*, **5**, 229.
CLEETON, G. U., and KNIGHT, F. B. (1924) : Validity of character judgments based on external criteria. *J. appl. Psychol.*, **8**, 215.
COLLINS, M. (1927) : British norms for the Pressey cross-out test. *Brit. J. Psychol.*, **18**, 121.
CONRAD, H. S. (1932) : The personal equation in ratings. I : An experimental determination. *J. genet. Psychol.*, **41**, 267.
—— (1933) : The personal equation in ratings. II : A systematic evaluation. *J. educ. Psychol.*, **24**, 39.
COPELAND, H. A. (1935) : A note on "The vectors of mind." *Psychol. Rev.*, **42**, 216.
COURTHIAL, A. (1931) : Emotional differences of delinquent and non-delinquent girls of normal intelligence. *Arch. Psychol.*, **20**, No. 133.
DEUTSCH, G. F. (1923) : *Conformity in Human Behavior*. Master's Thesis, Columbia University Library. Abstracted in Murphy (1937).
DEWAR, H. (1938) : A comparison of tests of artistic appreciation. *Brit. J. educ, Psychol.*, **8**, 29.
DOLL, E. A. (1936a) : *The Vineland Social Maturity Scale*. Vineland, New Jersey : The Training School.
—— (1936b) : Preliminary standardization of the Vineland social maturity scale. *Amer. J. Orthopsychiat.*, **6**, 283.
—— (1937) : A practical method for the measurement of social competence. *Eugen. Rev.*, **29**, 197.
DOWNEY, J. E. (1932) : Familial trends in personality. *Character & Pers.*, **1**, 35.
DROBA, D. D. (1932) : Methods for measuring attitudes. *Psychol. Bull.*, **29**, 309.
EAGLESHAM, E. J. R. (1937) : An enquiry concerning the practicability of typical educational aims. *Brit. J. educ. Psychol.*, **7**, 23.
EARLE, F. M., MILNER, M., et al (1929) : The use of performance tests of intelligence in vocational guidance. *Rep. industr. Fatig. Res. Bd., Lond.* No. **53**.
ELONEN, A. S., and WOODROW, H. (1928) : Group tests of psychopathic tendencies in children. *J. abnorm. soc. Psychol.*, **23**, 315.
ESTES, S. G. (1937) : *The Judgment of Personality on the Basis of Brief Records of Behavior*. Ph.D. Thesis, Harvard University Library
FARNSWORTH, P. R. (1937) : Changes in "attitude toward war" during the college years. *J. soc. Psychol.*, **8**, 274.
FERGUSON, L. W. (1935) : The influence of individual attitudes on construction of an attitude scale. *Ibid.*, **6**, 115.
FERNALD, G. G. (1912) : The defective delinquent class : Differentiating tests. *Amer. J. Insan.*, **68**, 523.

FISHER, V. E., and MARROW, A. J. (1934) : Experimental study of moods. *Character & Pers.*, **2**, 201.
FLANAGAN, J. C. (1935) : *Factor Analysis in the Study of Personality*. Stanford University Press.
FLEMMING, E. G. (1928) : The predictive value of certain tests of emotional stability as applied to college freshmen. *Arch. Psychol.*, **15**, No. 96.
FLUGEL, J. C., and RADCLYFFE, E. J. D. (1928) : The Pressey cross-out test compared with a questionnaire. *Brit. J. med. Psychol.*, **8**, 112.
FREYD, M. (1923) : The graphic rating scale. *J. educ. Psychol.*, **14**, 83.
—— (1924a) : Introverts and extroverts. *Psychol. Rev.*, **31**, 74.
—— (1924b) : The personalities of the socially and the mechanically inclined. *Psychol. Monogr.*, **33**, No. 151.
FRYER, D. (1931) : *The Measurement of Interests*. New York : Holt.
FURFEY, P. H. (1928) : Developmental age. *Amer. J. Psychiat.*, **8**, 149.
GARRETSON, O. K. (1930) : Relationships between expressed preferences and curricular abilities of ninth grade boys. *Teachers Coll. Contr. Educ., New York*. No. **396.**
GARRETT, H. E. (1924) : An empirical study of the various methods of combining incomplete order of merit ratings. *J. educ. Psychol.*, **15**, 157.
GILLILAND, A. R. (1926) : A revision and some results with the Moore-Gilliland aggressiveness test. *J. appl. Psychol.*, **10**, 143.
GOODENOUGH, F. L. (1928) : Measuring behavior traits by means of repeated short samples. *J. juv. Res.*, **12**, 230.
—— (1930) : Inter-relationships in the behavior of young children. *Child Develpm.*, **1**, 29.
GUILFORD, J. P. (1926) : An attempted study of emotional tendencies in criminals. *J. abnorm. soc. Psychol.*, **21**, 240.
—— (1936) : *Psychometric Methods*. New York : McGraw-Hill.
GUILFORD, J. P., and GUILFORD, R. B. (1934) : An analysis of the factors in a typical test of introversion-extroversion. *J. abnorm. soc. Psychol.*, **28**, 377.
—— —— (1936) : Personality factors S, E and M, and their measurement. *J. Psychol.*, **2**, 109.
GUNDLACH, R. H., and GERUM, E. (1931) : Vocational interests and types of ability. *J. educ. Psychol.*, **22**, 505.
GUTHRIE, E. R. (1927) : Measuring introversion and extroversion. *J. abnorm. soc. Psychol.*, **22**, 82.
HAGGERTY, M. E., OLSON, W. C., and WICKMAN, E. K. (1930) : *Behavior Rating Schedules*. Yonkers, New York : World Book Co.
HAMLEY, H. R., OLIVER, R. A. C., et al. (1937) : *The Educational Guidance of the School Child*. London : Evans.
HANNA, J. V. (1934) : Clinical procedure as a method of validating a measure of psychoneurotic tendency. *J. abnorm. soc. Psychol.*, **28**, 435.
HARPER, M. H. (1927) : Social beliefs and attitudes of American educators. *Teachers' Coll. Contr. Educ., New York*. No. **294.**
HARSH, C. M. (1936) : Three applications of cluster analysis to an annoyance study. *Psychol. Bull.*, **33**, 773.
HARTOG, P., RHODES, E. C., and BURT, C. (1936) : *The Marks of Examiners*. London : Macmillan.
HARTSHORNE, H., and MAY, M. A. (1928) : *Studies in Deceit*. New York : Macmillan.
HARVEY, O. L. (1932) : Concerning the Thurstone personality schedule. *J. soc. Psychol.*, **3**, 240.
HEIDBREDER, E. (1926) : Measuring introversion and extroversion. *J. abnorm. soc. Psychol.*, **21**, 120.
—— (1927) : The normal inferiority complex. *Ibid.*, **22**, 243.
HINCKLEY, E. D. (1932) : The influence of individual opinion on the construction of an attitude scale. *J. soc. Psychol.*, **3**, 283.
HOFFDITZ, E. L. (1934) : Family resemblances in personality traits. *Ibid.* **5**, 214.

HOLLINGWORTH, H. L. (1920) : *The Psychology of Functional Neuroses.* New York : Appleton.
—— (1929) : *Vocational Psychology and Character Analysis.* New York : Appleton.
HOLZINGER, K. J. (1936) : Recent research on unitary mental traits. *Character & Pers.*, **4**, 335.
—— (1937) : *Student Manual of Factor Analysis.* University of Chicago, Department of Education.
HORST, P. (1934) : Item analysis by the method of successive residuals. *J. exp. Educ.*, **2**, 254.
HOTELLING, H. (1933) : Analysis of a complex of statistical variables into principal components. *J. educ. Psychol.*, **24**, 417, 498.
HULL, C. L. (1928) : *Aptitude Testing.* Yonkers, New York : World Book Co.
HUMM, D. G., and WADSWORTH, G. W. (1935) : The Humm-Wadsworth temperament scale. *Amer. J. Psychiat.*, **92**, 163.
HUNT, J. M. (1936) : Psychological experiments with disordered persons. *Psychol. Bull.*, **33**, 1.
JASPER, H. H. (1930) : The measurement of depression-elation and its relation to a measure of extraversion-introversion. *J. abnorm. soc. Psychol.*, **25**, 307.
JOHNSON, W. B. (1934) : The effect of mood on personality traits as measured by Bernreuter. *J. soc. Psychol.*, **5**, 515.
JOHNSON, W. B., and TERMAN, L. M. (1935) : Personality characteristics of happily married, unhappily married, and divorced persons. *Character & Pers.*, **3**, 290.
JONES, M. C., and BURKS, B. S. (1936) : *Personality Development in Childhood.* Washington, D.C. : National Research Council.
JUNG, C. G. (1916) : *Collected Papers on Analytical Psychology* (trans. C. E. Long). London : Balliere, Tindall and Cox.
—— (1919) : *Studies in Word Association* (trans. M. D. Eder). London : Heinemann.
KATZ, D., and ALLPORT, F. H. (1931) : *Students' Attitudes.* Syracuse, New York : Craftsman Press.
KELLEY, E. L., MILES, C. C., and TERMAN, L. M. (1936) : Ability to influence one's score on a typical pencil-and-paper test of personality. *Character & Pers.*, **4**, 206.
KELLEY, T. L. (1923) : *Statistical Method.* New York : Macmillan.
—— (1928) : *Crossroads in the Mind of Man.* Stanford University Press.
—— (1935) : *Essential Traits of Mental Life.* Harvard University Press.
KELLEY, T. L., and KREY, A. C. (1934) : *Tests and Measurements in the Social Sciences.* New York : Scribner.
KENT, G. H., and ROSANOFF, A. J. (1910–11) : A study of association in insanity. *Amer. J. Insan.*, **67**, 37, 317.
KINGSBURY, F. A. (1922) : Analyzing ratings and training raters. *J. Person. Res.*, **1**, 377.
KIRKPATRICK, C. (1936) : Assumptions and methods in attitude measurements. *Amer. sociol. Rev.*, **1**, 75.
KIRKPATRICK, C., and STONE, S. (1935) : Attitude measurement and the comparison of generations. *J. appl. Psychol.* **19**, 564.
KLEIN, E. (1925) : *The Relation between One's Attitude to His Father and His Social Attitudes.* Master's Thesis, Columbia University Library. Abstracted in Murphy (1937).
KNIGHT, F. B. (1923) : The effect of the "acquaintance factor" upon personal judgments. *J. educ. Psychol.*, **14**, 129.
KULP, D. H. (1933) : The form of statements in attitude tests. *Sociol. & soc. Res.*, **18**, 18.
KULP, D. H., and DAVIDSON, H. H. (1934) : The application of the Spearman two-factor theory to social attitudes. *J. abnorm. soc. Psychol.*, **29**, 269.
LAIRD, D. A. (1925) : Detecting abnormal behaviour. *Ibid.*, **20**, 128.

LANDIS, C. (1925) : The justification of judgments. *J. Person. Res.*, **4**, 7.
—— (1932) : An attempt to measure emotional traits in juvenile delinquency. *Studies in the Dynamics of Behavior* (ed. K. S. Lashley). Chicago University Press. P. 263.
—— (1936) : Questionnaires and the study of personality. *J. nerv. ment. Dis.*, **83**, 125.
LANDIS, C., GULLETTE, R., and JACOBSEN, C. (1925) : Criteria of emotionality. *Pedagog. Sem.*, **32**, 209.
LANDIS, C., and PHELPS, L. W. (1928) : The prediction from photographs of success and of vocational aptitude. *J. exp. Psychol.*, **11**, 313.
LANDIS, C., ZUBIN, J., and KATZ, S. E. (1935) : Empirical evaluation of three personality adjustment inventories. *J. educ. Psychol.*, **26**, 321.
LASLETT, H. R., and BENNETT, E. (1934) : A comparison of scores on two measures of personality. *J. abnorm. soc. Psychol.*, **28**, 459.
LAYMAN, E. M. (1937) : An item analysis of the adjustment questionnaire. *Psychol. Bull.*, **34**, 782.
LEHMAN, H. C., and WITTY, P. A. (1927) : *The Psychology of Play Activities.* New York : Barnes.
LENTZ., T. F. (1930) : Utilizing opinion for character measurement. *J. soc. Psychol.*, **1**, 536.
—— (1934) : Reliability of opinionaire technique studied intensively by the retest method. *Ibid.*, **5**, 338.
LENTZ., T. F., HIRSHSTEIN, B., and FINCH, J. H. (1932) : Evaluation of methods of evaluating test items. *J. educ. Psychol.*, **23**, 344.
LICHTENSTEIN, A. (1934) : Can attitudes be taught ? *John Hopkins Univ. Stud. Educ., Baltimore, Md.* No. **21**.
LIKERT, R. (1932) : A technique for the measurement of attitudes. *Arch. Psychol.*, **22**, No. 140.
LIKERT, R., ROSLOW, S., and MURPHY, G. (1934) : A simple and reliable method of scoring the Thurstone attitude scales. *J. soc. Psychol.*, **5**, 228.
LURIE, W. A. (1937) : A study of Spranger's value-types by the method of factor analysis. *Ibid.*, **8**, 17.
MCCLATCHY, V. R. (1928) : A theoretical and statistical study of the personality trait originality as herein defined. *J. abnorm. soc. Psychol.*, **23**, 379.
MCCLOY, C. H. (1936) : A factor analysis of personality traits to underlie character education. *J. educ. Psychol.*, **27**, 375.
MCDONOUGH, M. R. (1929) : The empirical study of character. *Cath. Univ. Amer. Stud. Psychol. Psychiat., Washington, D.C.* **2**, Nos. 3 & 4.
MCGEOCH, J. A., and WHITELY, P. L. (1927) : The reliability of the Pressey X–O tests for investigating the emotions. *Pedagog. Sem.*, **34**, 255.
MAGSON, E. H. (1926) : How we judge intelligence. *Brit. J. Psychol. Monogr. Suppl.*, **3**, No. 9.
MALLER, J. B. (1932) : *Character Sketches.* Teachers' College, Columbia University : Bureau of Publications.
MANSON, G. E. (1931) : Occupational interests and personality requirements of women in business and the professions. *Mich. Bus. Stud.*, **3**, 281.
MATHEWS, C. O. (1927) : The effect of position of printed response words upon children's answers to questions in two-response types of tests. *J. educ. Psychol.*, **18**, 445.
MATHEWS, E. (1923) : A study of emotional instability in children. *J. Delinq.*, **8**, 1.
MAY, M. A., and HARTSHORNE, H. (1930) : Recent improvements in devices for rating character. *J. soc. Psychol.*, **1**, 66.
MELTZER, H. (1935) : Children's attitudes to parents. *Amer. J. Orthopsychiat.*, **5**, 244.
—— (1936) : Economic security and children's attitudes to parents. *Ibid.*, **6**, 590.
MILLER, L. W. (1934) : A critical analysis of the Peterson-Thurstone war attitude scale. *J. educ. Psychol.*, **25**, 662

MOORE, H. T. (1925): Innate factors in radicalism and conservatism. *J. abnorm. (soc.) Psychol.*, **20**, 234.
MORAN, T. F. (1935): A brief study of the validity of a neurotic inventory. *J. appl. Psychol.*, **19**, 180.
MORGAN, C. D., and MURRAY, H. A. (1935): A method for investigating fantasies. *Arch. Neurol. Psychiat.*, **34**, 289.
MURRAY, M. E. (1932): Validation of items of the psychoneurotic inventory. *J. Juv. Res.*, **16**, 213.
MURPHY, G. (1923): Types of word-association in dementia praecox, manic-depressives, and normal persons. *Amer. J. Psychiat.*, **2**, 539.
MURPHY, G., MURPHY, L. B., and NEWCOMB, T. M. (1937): *Experimental Social Psychology*. New York: Harper.
MYERS, C. S. (1932): Recent evidence of the value of vocational guidance. *Human Factor*, **6**, 438.
NEPRASH, J. A. (1936): The reliability of questions in the Thurstone personality schedule. *J. soc. Psychol.*, **7**, 239.
NEWCOMB, T. M. (1931): An experiment designed to test the validity of a rating technique. *J. educ. Psychol.*, **22**, 279.
NEYMANN, C. A., and KOHLSTEDT, K. D. (1929): A new diagnostic test for introversion-extroversion. *J. abnorm. soc. Psychol.*, **23**, 482.
O'CONNOR, J. (1928): *Born That Way*. Baltimore, Md.: Williams and Wilkins.
OLIVER, R. A. C. (1930): The traits of extroverts and introverts. *J. soc. Psychol.*, **1**, 345.
OLSON, W. C. (1929): The measurement of nervous habits in normal children. *Inst. Child Welfare Monogr. Ser., Minneapolis*. No. **3**.
—— (1930): *Problem Tendencies in Children*. University of Minnesota Press.
—— (1936): The waiver of signature in personal reports. *J. appl. Psychol.*, **20**, 442.
OLSON, W. C., and CUNNINGHAM, E. M. (1934): Time-sampling techniques. *Child Develpm.*, **5**, 41.
PAGE, J., LANDIS, C., and KATZ, S. E. (1934): Schizophrenic traits in the functional psychoses and in normal individuals. *Amer. J. Psychiat.*, **13**, 1213.
PARTEN, M. B. (1932): Social participation among pre-school children. *J. abn. soc. Psychol.*, **27**, 243.
PERRY, R. C. (1934): A group factor analysis of the adjustment questionnaire. *Southern Calif. Educ. Monogr., Los Angeles*. No. **5**.
PETERSON, R. C., and THURSTONE, L. L. (1933): *Motion Pictures and the Social Attitudes of Children*. New York: Macmillan.
PINTNER, R. (1918): Intelligence as estimated from photographs. *Psychol. Rev.*, **25**, 286.
—— (1933): A comparison of interests, abilities and attitudes. *J. abnorm. soc. Psychol.*, **27**, 351.
PRESSEY, S. L. (1921): A group scale for investigating the emotions. *Ibid.*, **16**, 55.
PRESSEY, S. L., and CHAMBERS, O. R. (1920): First revision of a group scale for investigating the emotions with tentative norms. *J. appl. Psychol.*, **4**, 97.
PRITCHARD, R. A. (1935): The relative popularity of secondary school subjects at various ages. *Brit. J. educ. Psychol.*, **5**, 157.
REMMERS, H. H. (1931): The equivalence of judgments to test items in the sense of the Spearman-Brown formula. *J. educ. Psychol.*, **21**, 66.
RICE, S. A. (1930): Statistical studies of social attitudes and public opinion. *Statistics in Social Studies*. University of Pennsylvania Press. P. 171.
RICHARDSON, M. W., and KUDER, G. F. (1933): Making a rating scale that measures. *Person. J.*, **12**, 36.
ROBINSON, C. E. (1937): Recent developments in the straw-poll field. *Publ. Opin. Quart.*, **1**, No. 3, 45.

RODGER, A. (1934) : Why and how the vocational psychologist studies temperament. *Human Factor*, **8**, 48.
ROSANDER, A. C. (1936) : The Spearman-Brown formula in attitude scale construction. *J. exp. Psychol.*, **19**, 486.
RUNDQUIST, E. A., and SLETTO, R. F. (1936) : *Personality in the Depression.* University of Minnesota Press.
SAFFIR, M. A. (1937) : A comparative study of scales constructed by three psychophysical methods. *Psychol. Bull.*, **34**, 716.
SCHWEGLER, R. A. (1929) : A study of introvert-extravert responses to certain test situations. *Teachers' Coll. Contr. Educ., New York.* No. **381.**
SCOTT, W. D., CLOTHIER, R. C., and MATHEWSON, S. B. (1923) : *Personnel Management.* Chicago : Shaw.
SHAKESPEARE, J. J. (1936) : An enquiry into the relative popularity of school subjects in elementary schools. *Brit. J. educ. Psychol.*, **6**, 147.
SHELLOW, S. M. (1931) : Vocational interest blank as an aid to interviewing. *Person. J.*, **9**, 379.
SHEN, E. (1925) : The influence of friendship upon personal ratings. *J. appl. Psychol.*, **9**, 66.
SIMPSON, R. M. (1934a) : A psychoneurotic inventory of penitentiary inmates. *J. soc. Psychol.*, **5**, 56.
—— (1934b) : Attitudes of teachers and prisoners toward seriousness of criminal acts. *J. crim. Law Criminol.*, **25**, 76.
SLETTO, R. F. (1936) : A critical study of the criterion of internal consistency in personality scale construction. *Amer. sociol. Rev.*, **1**, 61.
SMITH, R. B. (1932) : The development of an inventory for the measurement of inferiority feelings at the high school level. *Arch. Psychol.*, **22**, No. 144.
SPEARMAN, C. E. (1927) : *The Abilities of Man.* London : Macmillan.
SPEER, G. S. (1936) : The use of the Bernreuter personality inventory as an aid in the prediction of behavior problems. *J. juv. Res.*, **20**, 65.
SPRANGER, E. (1928) : *Types of Men* (trans. P. J. W. Pigors). Halle : Niemeyer.
STAGNER, R. (1933) : Methodology of attitude measurement. *Research in Social Psychology of Rural Life* (ed. J. D. Black). New York : Social Science Research Council, Bull. No. **17.**
—— (1934) : Validity and reliability of the Bernreuter personality inventory. *J. abnorm. soc. Psychol.*, **28**, 413.
—— (1937) : The Wisconsin scale of personality traits. *Ibid.*, **31**, 463.
STEINMETZ, H. C. (1932) : Measuring ability to fake occupational interest. *J. appl. Psychol.*, **16**, 123.
STEPHENSON, W. (1936a) : The inverted factor technique. *Brit. J. Psychol.*, **26**, 344.
—— (1936b) : Introduction to inverted factor analysis, with some applications to studies in orexis. *J. educ. Psychol.*, **27**, 353.
—— (1936c) : Some recent contributions to the theory of psychometry. *Character & Pers.*, **4**, 294.
—— (1936d) : A new application of correlation to averages. *Brit. J. educ. Psychol.*, **6**, 43.
STOGDILL, R. M. (1934) : Attitudes of parents toward parental behavior. *J. abnorm. soc. Psychol.*, **29**, 293.
STOUFFER, S. A. (1931) : Experimental comparison of a statistical and a case history technique of attitude research. *Pub. Amer. sociol. Soc.*, **25**, 154.
STRONG, E. K. (1927) : *Vocational Interest Blank.* Stanford University Press.
—— (1931) : *Change of Interests with Age.* Stanford University Press.
—— (1934) : Classification of occupations by interests. *Person. J.*, **12**, 301.
—— (1935) : Predictive value of the vocational interest test. *J. educ. Psychol.*, **26**, 331.
STUTSMAN, R. (1931) : *Mental Measurement of Pre-school Children.* Yonkers, New York : World Book Co.

SYMONDS, P. M. (1924) : On the loss of reliability in ratings due to coarseness of the scale. *J. exp. Psychol.*, **7**, 456.
—— (1931) : *Diagnosing Personality and Conduct.* New York : Century.
—— (1934) : *Psychological Diagnosis in Social Adjustment.* New York : American Book Co.
—— (1936) : Influence of order of presentation of items in ranking. *J. educ. Psychol.*, **27**, 445.
TAYLOR, H. C. (1934) : Social agreement on personality traits as judged from speech. *J. soc. Psychol.*, **5**, 244.
TELFORD, C. W. (1934) : An experimental study of some factors influencing the social attitudes of college students. *Ibid.*, **5**, 421.
TERMAN, L. M. (1925) : *Genetic Studies of Genius.* Vol. I. *Mental and Physical Traits of* 1,000 *Gifted Children.* Stanford University Press.
TERMAN, L. M., and MILES, C. C. (1936) : *Sex and Personality.* New York : McGraw Hill.
THOMAS, D. S., LOOMIS, A. M., and ARRINGTON, R. E. (1933) : *Observational Studies of Social Behavior.* Yale University, Institute of Human Relations.
THOMPSON, L. A., and REMMERS, H. H. (1928) : Some observations concerning the reliability of the Pressey X–O test. *J. appl. Psychol.*, **12**, 477.
THOMSON, G. H. (1935) : On complete families of correlation coefficients, and their tendency to zero tetrad-differences. *Brit. J. Psychol.*, **26**, 63.
THORNDIKE, E. L. (1920) : A constant error in psychological ratings. *J. appl. Psychol.*, **4**, 25.
—— (1935) : The interests of adults. *J. educ. Psychol.*, **26**, 401, 497.
—— (1936) : The value of reported likes and dislikes for various experiences and activities as indications of personal traits. *J. appl. Psychol.*, **20**, 285.
THORNDIKE, E. L., et al. (1926) : *The Measurement of Intelligence.* Teachers' College, Columbia University : Bureau of Publications.
THOULESS, R. H. (1935) : The tendency to certainty in religious belief. *Brit. J. Psychol.*, **26**, 16.
—— (1936) : Test unreliability and function fluctuation. *Ibid.*, **26**, 325.
THURSTONE, L. L. (1927a) : A law of comparative judgment. *Psychol. Rev.*, **34**, 273.
—— (1927b) : The method of paired comparisons for social values. *J. abnorm. soc. Psychol.*, **21**, 384.
—— (1928) : An experimental study of nationality preferences. *J. gen. Psychol.*, **1**, 405.
—— (1929) : Theory of attitude measurement. *Psychol. Rev.*, **36**, 222.
—— (1930) : A scale for measuring attitude toward the movies. *J. educ. Res.*, **22**, 89.
—— (1931a) : Influence of motion pictures on children's attitudes. *J. soc. Psychol.*, **2**, 291.
—— (1931b) : A multiple factor study of vocational interests. *Person. J.*, **10**, 198.
—— (1933) : *A Simplified Multiple Factor Method.* University of Chicago Bookstore.
—— (1934) : The vectors of mind. *Psychol. Rev.*, **41**, 1.
—— (1935) : *The Vectors of Mind.* University of Chicago Press.
THURSTONE, L. L., and CHAVE, E. J. (1929) : *The Measurement of Attitude.* University of Chicago Press.
THURSTONE, L. L., and THURSTONE, T. G. (1930) : A neurotic inventory. *J. soc. Psychol.*, **1**, 3.
TJADEN, J. C. (1926) : Emotional reactions of delinquent boys of superior intelligence compared with those of college students. *J. abnorm. soc. Psychol.*, **21**, 192.
TRYON, C. M. (1933) : A study of social and emotional adjustments. Ph.D. Thesis. University of California Library. Abstracted in Jones and Burks (1936).

UEHLING, H. F. (1934) : Comments on a suggested revision of the Woodworth psychoneurotic inventory. *J. abnorm. soc. Psychol.*, **28,** 462.
UHRBROCK, R. S. (1930) : Estimating intelligence from photographs. *Proc. IX Int. Cong. Psychol.* Princeton, New Jersey : Psychological Review Co. P. 451.
—— (1932) : Rating tendencies of personally selected judges. *J. educ. Psychol.*, **23,** 594.
—— (1934) : Attitudes of 4,430 employees. *J. soc. Psychol.*, **5,** 365.
VALENTINE, C. W. (1934) : An enquiry as to reasons for the choice of the teaching profession by University students. *Brit. J. educ. Psychol.*, **4,** 237.
VERNON, P. E. (1929) : Tests of temperament and personality. *Brit. J. Psychol.*, **20,** 97.
—— (1933a) : Some characteristics of the good judge of personality. *J. soc. Psychol.*, **4,** 42.
—— (1933b) : The American *v.* the German methods of approach to the study of temperament and personality. *Brit. J. Psychol.*, **24,** 156.
—— (1934a) : The measurement of personality and temperament. *Human Factor*, **8,** 87.
—— (1934b) : The attitude of the subject in personality testing. *J. appl. Psychol.*, **18,** 165.
—— (1935) : Tests in aesthetics. *The Testing of Intelligence* (ed. H. R. Hamley). London : Evans. P. 133.
—— (1936) : The matching method applied to investigations of personality. *Psychol. Bull.*, **33,** 149.
—— (1938) : Questionnaires, attitude tests and rating scales. *Problems and Methods in Social Psychology and Sociology* (ed. E. J. Lindgren).
VERNON, P. E., and ALLPORT, G. W. (1931) : A test for personal values. *J. abnorm. soc. Psychol.*, **26,** 231.
VETTER, G. B. (1930) : The measurement of social and political attitudes and the related personality factors. *Ibid.*, **25,** 149.
WANG, C. K. A. (1932a) : A scale for measuring persistence. *J. soc. Psychol.*, **3,** 79.
—— (1932b) : Suggested criteria for writing attitude statements. *Ibid.*, **3,** 367.
WATSON, G. B. (1925) : The measurement of fairmindedness. *Teachers' Coll. Contr. Educ., New York.* No. **176.**
—— (1929) : Orient and Occident : An opinion study. *Relig. Educ.*, **24,** 322.
—— (1934) : A comparison of the effects of lax versus strict home training. *J. soc. Psychol.*, **5,** 102.
WEBB, E. (1915) : Character and intelligence. *Brit. J. Psychol. Monogr. Suppl.*, **1,** No. 3.
WEBER, C. O. (1932) : Further tests of the Wells emotional age scale. *J. abnorm. soc. Psychol.*, **27,** 65.
WEBER, C. O., and MAIJGREN, R. (1929) : The experimental differentia of introversion and extraversion. *J. genet. Psychol.*, **36,** 571.
WELLS, F. L. (1919) : Association type and personality. *Psychol. Rev.*, **26,** 371.
WHEAT, L. B. (1931) : Free associations to common words. *Teachers' Coll. Contr. Educ., New York.* No. **498.**
WHISLER, L. D. (1934) : Multiple-factor analysis of generalized attitudes. *J. soc. Psychol.*, **5,** 283.
WICKMAN, E. K. (1928) : *Children's Behavior and Teachers' Attitudes.* New York : Commonwealth Fund, Division of Publications.
WILLOUGHBY, R. R. (1932a) : A scale of emotional maturity. *J. soc. Psychol.*, **3,** 3.
—— (1932b) : Some properties of the Thurstone personality schedule and a suggested revision. *Ibid.*, **3,** 401.
—— (1934–36) : Neuroticism in Marriage. *Ibid.*, **5,** 3, 467 ; **6,** 397 ; **7,** 19.
—— (1937) : The emotionality of spinsters. *Character & Pers.*, **5,** 215.

WILLOUGHBY, R. R., and MORSE, M. E. (1936) : Spontaneous reactions to a personality inventory. *Amer. J. Orthopsychiat.*, **6**, 562.
WOLF, R., and MURRAY, H. A. (1936) : An experiment in judging personalities. *J. Psychol.*, **3**, 345.
WOODROW, H., and LOWELL, F. (1916) : Children's association frequency tables. *Psychol. Rev. Monogr.*, **22**, No. 97.
WOODWORTH, R. S. (1920) : *Personal Data Sheet.* Chicago : Stoelting.
WYATT, S., and LANGDON, J. N. (1937) : Fatigue and boredom in repetitive work. *Rep. industr. Hlth. Res. Bd., Lond.* No. **77**.
WYMAN, J. B. (1925) : Interest tests of a group of gifted children. *Genetic Studies of Genius*, Vol. I (ed. L. M. Terman). Stanford University Press. P. 455.
ZIMMERMAN, F. K. (1934) : Religion a conservative social force. *J. abnorm. soc. Psychol.*, **28**, 473.
ZUBIN, J. (1934) : The method of internal consistency for selecting test items. *J. educ. Psychol.*, **25**, 345.

INDEX OF THE MAIN TESTS, METHODS AND INVESTIGATIONS

References are to paragraphs, not pages.

REPORTS OF THE
INDUSTRIAL HEALTH RESEARCH BOARD

Only those Reports marked with an asterisk are available at present and may be obtained as indicated on cover page iv. Prices in brackets include postage.
It is anticipated that other Reports will be reprinted in due course.

No. 1.—The Influence of Hours of Work and of Ventilation on Output in Tinplate Manufacture, by H. M. Vernon. (1919.)

No. 2.—The Output of Women Workers in relation to Hours of Work in Shell-making, by Ethel E. Osborne. (1919.)

No. 3.—A Study of Improved Methods in an Iron Foundry, by C. S. Myers. (1919.)

No. 4.—The Incidence of Industrial Accidents upon Individuals, with special reference to Multiple Accidents, by M. Greenwood and Hilda M. Woods. (1919.)

No. 5.—Fatigue and Efficiency in the Iron and Steel Industry, by H. M. Vernon. (1920.)

No. 6.—The Speed of Adaptation of Output to altered Hours of Work, by H. M. Vernon. (1920.)

No. 7.—Individual Differences in Output in the Cotton Industry, by S. Wyatt. (1920.)

No. 8.—Some Observations on Bobbin Winding, by S. Wyatt and H. C. Weston. (1920.)

No. 9.—A Study of Output in Silk Weaving during the Winter Months, by P. M. Elton. (1920.)

No. 10.—Preliminary Notes on the Boot and Shoe Industry, by J. Loveday and S. H. Munro. (1920.)

No. 11.—Preliminary Notes on Atmospheric Conditions in Boot and Shoe Factories, by W. D. Hambly and T. Bedford (1921)

No. 12.—Vocational Guidance (a Review of the Literature), by B. Muscio. (1921.)

No. 13.—A Statistical Study of Labour Turnover in Munition and other Factories, by G. M. Broughton, E. M. Newbold and E. C. Allen. (1921.)

No. 14.—Time and Motion Study, by E. Farmer. (1921.)

No. 15.—Motion Study in Metal Polishing, by E. Farmer, assisted by R. S. Brooke. (1921.)

No. 16.—Three Studies in Vocational Selection, by B. Muscio and E. Farmer, assisted by A. B. B. Eyre. (1922.)

No. 17.—An Analysis of the Individual Differences in the Output of Silk-Weavers, by P. M. Elton. (1922)

No. 18.—Two Investigations in Potters' Shops, by H. M. Vernon and T. Bedford. (1922.)

No. 41.—Rest Pauses in Heavy and Moderately Heavy Industrial Work, by H. M. Vernon and T. Bedford, assisted by C. G. Warner. (1927.)

No. 42.—Rest Pauses in Industry (a Review of the Results Obtained), by S. Wyatt. (1927.)

No. 43.—A Study of Telegraphists' Cramp, by May Smith, Millais Culpin and Eric Farmer. (1927.)

No. 44.—The Physique of Women in Industry (a Contribution towards the Determination of the Optimum Load), by E. P. Cathcart, E. M. Bedale, C. Blair, K. Macleod and E. Weatherhead, with a special section by Sybil G. Overton. (1927.)

No. 45.—Two Contributions to the Experimental Study of the Menstrual Cycle: (I)—Its Influence on Mental and Muscular Efficiency, by S. C. M. Sowton and C. S. Myers. (II)—Its Relation to General Functional Activity, by E. M. Bedale. (1928.)

No. 46.—A Physiological Investigation of the Radiant Heating in Various Buildings, by H. M. Vernon and M. D. Vernon, assisted by Isabel Lorrain-Smith. (1928.)

No. 47.—Two Studies on Hours of Work. (I)—Five-Hour Spells for Women, with reference to Rest Pauses, by H. M. Vernon and M. D. Vernon, assisted by I. Lorrain-Smith. (II)—The Two-Shift System in Certain Factories, by May Smith and M. D. Vernon. (1928.)

No. 48.—Artificial Humidification in the Cotton Weaving Industry. Its Effect upon the Sickness Rates of Weaving Operatives, by A. Bradford Hill. (1927.)

No. 49.—On the Relief of Eyestrain among Persons Performing Very Fine Work, by H. C. Weston and S. Adams. (1928.)

No. 50.—The Physiological Cost of the Muscular Movements involved in Barrow Work, by G. P. Crowden. (1928.)

No. 51.—A Study of Absenteeism in a Group of Ten Collieries, by H. M. Vernon and T. Bedford, assisted by C. G. Warner. (1928.)

No. 52.—The Comparative Effects of Variety and Uniformity in Work, by S. Wyatt and J. A. Fraser, assisted by F. G. L. Stock. (1928.)

No. 53.—The Use of Performance Tests of Intelligence in Vocational Guidance, by F. M. Earle, M. Milner and others. (1929.)

No. 54.—An Investigation into the Sickness Experience of Printers (with special reference to the Incidence of Tuberculosis), by A. Bradford Hill. (1929.)

No. 55.—A Study of Personal Qualities in Accident Proneness and Proficiency, by Eric Farmer and E. G. Chambers. (1929.)

No. 56.—The Effects of Monotony in Work: A Preliminary Enquiry, by S. Wyatt and J. A. Fraser, assisted by F. G. L. Stock. (1929.)

No. 57.—Further Experiments on the Use of Special Spectacles in Very Fine Processes, by H. C. Weston and S. Adams. (1929.)

No. 58.—A Study of Heating and Ventilation in Schools, by H. M. Vernon and T. Bedford, assisted by C. G. Warner. (1930.)

No. 19.—Two Contributions to the Study of Accident Causation, by Ethel E. Osborne, H. M. Vernon and B. Muscio. (1922.)

No. 20.—A Study of Efficiency in Fine Linen Weaving, by H. C. Weston. (1922.)

No. 21.—Atmospheric Conditions in Cotton Weaving, by S. Wyatt. (1923.)

No. 22.—Some Studies in the Laundry Trade, by May Smith. (1922.)

No. 23.—Variations in Efficiency in Cotton Weaving, by S. Wyatt. (1923.)

No. 24.—A Comparison of Different Shift Systems in the Glass Trade, by E. Farmer, assisted by R. S. Brooke and E. G. Chambers. (1923.)

No. 25.—Two Studies on Rest Pauses in Industry, by H. M. Vernon and T. Bedford, assisted by C. G. Warner. (1924.)

No. 26.—On the Extent and Effects of Variety in Repetitive Work, by H. M. Vernon, S. Wyatt and A. D. Ogden. (1924.)

*No. 27.—Results of Investigation in Certain Industries. (1924.) 6*d.* (7½*d.*)

No. 28.—The Function of Statistical Method in Scientific Investigation, by G. Udny Yule. (1924.)

No. 29.—The Effects of Posture and Rest in Muscular Work, by E. M. Bedale and H. M. Vernon. (1924.)

No. 30.—An Experimental Investigation into Repetitive Work, by Isabel Burnett. (1925.)

No. 31.—Performance Tests of Intelligence, by Frances Gaw. (1925.)

No. 32.—Studies in Repetitive Work, with special reference to Rest Pauses, by S. Wyatt, assisted by J. A. Fraser. (1925.)

No. 33.—A Study in Vocational Guidance by Frances Gaw, Lettice Ramsey, May Smith and Winifred Spielman under the general direction of Cyril Burt. (1926.)

*No. 34.—A Contribution to the Study of the Human Factor in the Causation of Accidents, by E. M. Newbold. (1926.) 7*s.* 6*d.* (7*s.* 10*d.*)

No. 35.—A Physiological Study of the Ventilation and Heating in Certain Factories, by H. M. Vernon and T. Bedford, assisted by C. G. Warner. (1926.)

No. 36.—On the Design of Machinery in relation to the Operator, by L. A. Legros and H. C. Weston. (1926.)

No. 37.—Fan Ventilation in a Humid Weaving Shed. An experiment carried out for the Departmental Committee on Humidity in Cotton Weaving, by S. Wyatt, assisted by J. A. Fraser and F. G. L. Stock. (1926.)

No. 38.—A Psychological Study of Individual Differences in Accident Rates, by E. Farmer and E. G. Chambers. (1926.)

No. 39.—The Relation of Atmospheric Conditions to the Working Capacity and the Accident Rate of Coal Miners, by H. M. Vernon and T. Bedford, assisted by C. G. Warner. (1927.)

No. 40.—The Effect of Eyestrain on the Output of Linkers in the Hosiery Industry, by H. C. Weston and S. Adams. (1927.)

No. 59.—Sickness amongst Operatives in Lancashire Cotton Spinning Mills (with special reference to the Cardroom), by A. Bradford Hill. (1930.)

No. 60.—The Atmospheric Conditions in Pithead Baths, by H. M. Vernon and T. Bedford, assisted by C. G. Warner. (1930.)

*No. 61.—The Nervous Temperament, by Millais Culpin and May Smith. (1930.) 4*s.* 6*d.* (4*s.* 7½*d.*)

No. 62.—Two Studies of Absenteeism in Coal Mines. (I)—The Absenteeism of Miners in Relation to Short Time and other Conditions, by H. M. Vernon and T. Bedford, assisted by C. G. Warner. (II)—A Study of Absenteeism at certain Scottish Collieries, by T. Bedford and C. G. Warner. With Appendix by E. P. Cathcart and James Taylor. (1931.)

No. 63.—Inspection Processes in Industry: a Preliminary Report, by S. Wyatt and J. N. Langdon. (1932.)

No. 64.—A Classification of Vocational Tests of Dexterity, by A. E. Weiss Long and T. H. Pear. (1932.)

No. 65.—Two Studies in the Psychological Effects of Noise. (I)—Psychological Experiments on the Effects of Noise, by K. G. Pollock and F. C. Bartlett. (II)—The Effects of Noise on the Performance of Weavers, by H. C. Weston and S. Adams. (1932.)

No. 66.—An Experimental Study of Certain Forms of Manual Dexterity by J. N. Langdon. (1932.)

No. 67.—Manual Dexterity: Effects of Training. (I)—Transfer of Training in Manual Dexterity and Visual Discrimination, by E. M. Henshaw, P. Holman and J. N. Langdon. (II)—Distribution of Practice in Manual Dexterity, by E. M. Henshaw and P. Holman. (1933.)

No. 68.—Tests for Accident Proneness, by E. Farmer, E. G. Chambers and F. J. Kirk. (1933.)

No. 69.—Incentives in Repetitive Work. A Practical Experiment in a Factory, by S. Wyatt, assisted by L. Frost and F. G. L. Stock. (1934.)

No. 70.—The Performance of Weavers under Varying Conditions of Noise, by H. C. Weston and S. Adams. (1935.)

No. 71.—The Physique of Man in Industry, by E. P. Cathcart, D. E. R. Hughes and J. G. Chalmers. (1935.)

*No. 72.—Incentives. Some Experimental Studies, by C. A. Mace. (1935.) 5*s.* 6*d.* (5*s.* 9*d.*)

No. 73.—The Acquisition of Skill: an Analysis of Learning Curves, by J. M. Blackburn. (1936.)

No. 74.—The Prognostic Value of some Psychological Tests, by E. Farmer and E. G. Chambers. (1936.)

*No. 75.—Sickness Absence and Labour Wastage. Part I, by May Smith and Margaret A. Leiper. Part II, by Major Greenwood and May Smith. (1936.) 6*s.* (6*s.* 3*d.*)

No. 76.—The Warmth Factor in Comfort at Work. A Physiological Study of Heating and Ventilation, by T. Bedford. (1936.)

*No. 77.—Fatigue and Boredom in Repetitive Work, by S. Wyatt and J. N. Langdon, assisted by F. G. L. Stock. (1937.) 6*s.* 6*d.* (6*s.* 8*d.*)

No. 78.—A Borstal Experiment in Vocational Guidance, by Alec Rodger. (1937.)

No. 79.—An Investigation into the Sickness Experience of London Transport Workers, with special reference to Digestive Disturbances, by A. Bradford Hill. (1937.)

No. 80.—Toxicity of Industrial Organic Solvents: Summaries of Published Work, Compiled by Ethel Browning under the direction of the Committee on the Toxicity of Industrial Solvents. (1937.) *(Under revision)*

No. 81.—The Effects of Conditions of Artificial Lighting on the Performance of Worsted Weavers, by H. C. Weston. (1938.)

No. 82.—The Machine and the Worker: a Study of Machine-Feeding Processes, by S. Wyatt and J. N. Langdon, assisted by F. G L. Stock. (1938.)

*No. 83.—The Assessment of Psychological Qualities by Verbal Methods: a Survey of Attitude Tests, Rating Scales and Personality Questionnaires, by P. E. Vernon. (1938.) 8*s.* 6*d.* (8*s.* 9*d.*)

No. 84.—A Study of Accident Proneness amongst Motor Drivers, by E. Farmer and E. G. Chambers. (1940.)

*No. 85.—The Recording of Sickness Absence in Industry (A Preliminary Report), by a Sub-Committee of the Industrial Health Research Board. (1944.) (Reprinted 1948.) 6*d.* (7½*d.*)

*No. 86.—A Study of Certified Sickness Absence among Women in Industry, by S. Wyatt, assisted by R. Marriott, W. M. Dawson, Norah M. Davis, D. E. R. Hughes and F. G. L. Stock. (1945.) 9*d.* (10½*d.*)

No. 87.—The Relation between Illumination and Visual Efficiency—the Effect of Brightness Contrast, by H. C. Weston. (1945.)

No. 88.—A Study of Women on War Work in Four Factories, by S. Wyatt, assisted by R. Marriott, W. M. Dawson, Norah M. Davis, D. E. R. Hughes and F. G. L. Stock. (1945.)

*No. 89.—Artificial Sunlight Treatment in Industry. A Report on the Results of Three Trials—in an Office, a Factory and a Coalmine, by Dora Colebrook. (1946.) 1*s.* (1*s.* 1½*d.*)

*No. 90.—The Incidence of Neurosis among Factory Workers, by Russell Fraser, with the collaboration of Elizabeth Bunbury, Barbara Danniell, M. Elizabeth Barling, F. Estelle Waldron, P. Mary Kemp and Imogen Le[illegible]. (1947.) 1*s.* 3*d.* (1*s.* 5*d.*)

REPORTS CLASSIFIED ACCORDING TO SUBJECT MATTER.

1. *Hours of work, rest pauses, etc.* *Nos.* 1, 2, 5, 6, 24, 41, 47, 56, 88.
2. *Dexterity.* *Nos.* 63, 64, 66, 67, 73.
3. *Industrial accidents.* *Nos.* 4, 19, 39, 55, 62, 68, 84.
4. *Atmospheric conditions.* *Nos.* 1, 5, 11, 18, 20, 21, 22, 23, 37, 39, 46, 48, 51, 58, 59, 60, 76.
5. *Vision and lighting.* *Nos.* 9, 20, 23, 40, 49, 57, 81, 87.
6. *Vocational guidance, selection.* *Nos.* 12, 16, 31, 33, 43, 53, 55, 61, 64, 74, 78, 83.
7. *Time and motion study, methods of work.* *Nos.* 3, 7, 8, 9, 14, 15, 17, 22, 23, 30, 52, 56, 72, 77, 82.
8. *Posture and physique.* *Nos.* 15, 16, 29, 36, 44, 50, 71.
9. *Sickness and absenteeism.* *Nos.* 51, 54, 62, 75, 79, 85, 86, 88, 89.
10. *Noise.* *Nos.* 65, 70.
11. *Toxicity hazards.* *No.* 80.
12. *Women in industry.* *Nos.* 2, 44, 45, 86, 88.
13. *Neurosis.* *Nos.* 61, 90

REPORTS CLASSIFIED ACCORDING TO INDUSTRY

(*a*) *Mining.* *Nos.* 39, 51, 60, 62.
(*b*) *Metals and engineering* *Nos.* 1, 2, 3, 5, 6, 15.
(*c*) *Textiles.* *Nos.* 7, 8, 9, 17, 20, 21, 23, 37, 40, 48, 49, 59, 70, 81.
(*d*) *Boots and shoes.* *Nos.* 10, 11.
(*e*) *Pottery.* *No.* 13.
(*f*) *Laundry.* *No.* 22.
(*g*) *Glass.* *No.* 24.
(*h*) *Printing.* *Nos.* 16, 26, 54.
(*i*) *Transport.* *Nos.* 79, 84.
(*j*) *Light repetitive work.* *Nos.* 14, 25, 26, 30, 32, 52.
(*k*) *Muscular work.* *Nos.* 29, 41, 44, 50, 71.